12-2018

BAYESIAN INTEGRATIVE ANALYSIS OF OMICS DATA

Youyi Zhang

BAYESIAN INTEGRATIVE ANALYSIS OF OMICS DATA

A

DISSERTATION

Presented to the Faculty of

The University of Texas

MD Anderson Cancer Center UTHealth

Graduate School of Biomedical Sciences

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

by

Youyi Zhang, M.S.

Houston, Texas

December, 2018

BAYESIAN INTEGRATIVE ANALYSIS OF OMICS DATA

Youyi Zhang, M.S

Advisory Professor: Veerabhadran Baladandayuthapani, Ph.D.

Jeffrey S. Morris, Ph.D.

Technological innovations have produced large multi-modal datasets that range in multiplatform genomic data, pathway data, proteomic data, imaging data and clinical data. Integrative analysis of such data sets have potentiality in revealing important biological and clinical insights into complex diseases like cancer. This dissertation focuses on Bayesian methodology establishment in integrative analysis of radiogenomics and pathway driver detection applied in cancer applications. We initially present Radio-iBAG that utilizes Bayesian approaches in analyzing radiological imaging and multi-platform genomic data, which we establish a multi-scale Bayesian hierarchical model that simultaneously identifies genomic and radiomic, i.e., radiology-based imaging markers, along with the latent associations between these two modalities, and to detect the overall prognostic relevance of the combined markers. Our method is motivated by and applied to The Cancer Genome Atlas glioblastoma multiforme data set, wherein it identifies important magnetic resonance imaging features and the associated genomic platforms that are also significantly related with patient survival times. For another aspect of integrative analysis, we then present pathDrive that aims to detect key genetic and epigenetic upstream drivers that influence pathway activity. The method is applied into colorectal cancer incorporated with its four molecular subtypes. For each of the pathways that significantly differentiates subgroups, we detect important genomic drivers that can be viewed as “switches” for the pathway activity. To extend the analysis, finally, we develop proteomic based pathway driver analysis for multiple cancer types wherein we simultaneously detect genomic upstream factors that influence a specific pathway for each cancer type within the cancer group. With Bayesian hierarchical model, we detect signals borrowing strength from common cancer type to rare cancer type, and

simultaneously estimate their selection similarity. Through simulation study, our method is demonstrated in providing many advantages, including increased power and lower false discovery rates. We then apply the method into the analysis of multiple cancer groups, wherein we detect key genomic upstream drivers with proper biological interpretation. The overall framework and methodologies established in this dissertation illustrate further investigation in the field of integrative analysis of omics data, provide more comprehensive insight into biological mechanisms and processes, cancer development and progression.

Acknowledgements

First and foremost, I would like convey my great gratitude to my research advisors Dr. Veera Baladandayuthapani and Dr. Jeffrey Morris for their great support and guidance throughout the entire period of my graduate work. They have provided me with constructive and thoughtful suggestions which have immeasurable contributions to my professional growth. I feel extremely grateful for their consistent instruction and encouragement.

I would also like to thank my dissertation comittee members, Dr. Sadhan Majumder, Dr. Christine Peterson, Dr. Rehan Akbani and Dr. Dawid Schellingerhout for their direction and invaluable suggestions ranging from genetics study, Bayesian statistics, computational biology and radiology perspectives on my research projects. I deeply appreciate their time and helpful comments.

I would like to thank Dr. Gani Manyam for his bioinformatics expertise, his contribution to our publication, and his great patience on my frequent data-related questions.

I would like to express my appreciation to Dr. Rao and Shivali Narang Aerry for their radiomics expertise, collaboration to our publication, and their offering of the wonderful resources for my study.

I would also like to acknowledge Dr. Loki Natarajan for introducing me to the field of biostatistics and providing me with constant encouragement throughout these years. For her great support, I will be forever grateful.

I would like to acknowledge my friends and peers who provided me with great help and support, you make my journey more joyful.

Finally, I would like to thank Quantitative Science program of MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences for its well-established curriculum, sufficient training opportunities and great doctorate educational structure.

To my beloved parents, without whose never-failing support and encouragement, I could not have made it this far.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

High-throughput technologies increasingly enable the advent of probing multiple biological layers in parallel, which involves quantitative monitor of multiple profiles ranging from genome, transcriptome, epigenome, proteome to phenome profiling [39]. This comprehensive assessment of a molecular set is termed "omics", with further extension as "multi-omics" referring to biological analysis methods where the data sets contain multiple "omes". In order to achieve more comprehensive investigation into biological processes, integrative analyses that utilize information across these multiple data modalities have become a promising biological research area. The topic of integrative analysis of multi-omics data faces several main challenges: the complicated biological systems and processes, limitation in measuring technologies, high dimensionality with large number of variables and relatively lower sample size. To address the challenges, different types of integrative approaches applied to a wide range of biological problems have been established.

From methodological perspective, to classify the existing methods, the typical criterion is whether it belongs to bayesian (BY) approach [12] [99] where it incorporates prior information into the modeling scheme. From application perspective, one type

of research involves the study in molecular biological regulatory mechanisms, which unravel complex cellular biological processes. This type of studies has been shown to have promising advantages, such as higher statistical power and decreased false discovery rates as Wang et. al., 2013 [112] illustrates. Besides integrative analysis at molecular level, another research area is the study of Radiomics data which refers to the high-throughput extraction of numbers of image features from images [58]. Furthermore, a recent research area termed "radiogenomics" which creates a link between molecular properties with imaging phenotypes has drawn great attention of researchers. The research field incorporates novel approaches digging into the hidden associations between gene expression patterns with radiomic phenotypes [94] [45].

Another integrative analyzing application area is related with Network Biology which has become a nascent and burgeoning subfield of systems biology that involves the discovery and characterization of molecular interactions underlying complex diseases, including cancer. One aspect of this field is the study of molecular pathways that have been discovered and curated by systems biologists. Different methods have been developed to accurately measure pathway activity, such as Gene set enrichment analysis (GSEA) [104] and pathway level gene expression analysis using singular value decomposition [109], which all alleviate the complexity in analyzing large amount of individual genes or proteins and simultaneously provide clearer picture of biological functional processes, components or structures [54].

Integrative analysis for one particular cancer type has been widely applied such as Hu et al., 2017 [45] which conducts radiogenomic analysis into the investigation of genetic heterogeneity in glioblastoma. Current integrative analysis has extended to the analysis of pan-cancer which initiatively targets to examine the similarities and discrepancies among the genomics, pathway or cellular functions across multiple

tumor types [114]. This kind of analysis gives insights on how one type of cancer associates to another cancer types at genomic level, epigenomic level, transcriptomic, pathway level and clinical level.

Following the literature review described above, this dissertation focuses on integrative analysis of radiogenomics, pathway genetic and epigenetic driver detection applied in single cancer type and multiple cancer types, with the purpose of identifying genomic and radiomic targets that significantly related with clinical outcomes, target switches that drive pathway activity in one cancer and in pan-cancer applications. We mainly apply Bayesian methods for the integrative analysis where we incorporate prior settings into the analyzing framework taking account of multi-scale data sets, high dimensionality and similarity in the patterns of key factors across multiple cancer types.

More specifically, in Chapter 2, motivated by "Bayesian methods for expression-based integration of various types of genomics data" that was established by Jennings et al., 2013 [49] which investigates clinical related genomic markers taking account of molecular regulatory mechanisms. In this chapter, we develop a more in-depth illustration of Radio-iBAG which refers to Radiomics-based Integrative Bayesian Analysis of Multiplatform Genomic Data that further relates genomics with radiomics and clinical outcomes. We apply our methodology framework to the case study of glioblastoma multiforme (GBM), where we detect key genomic and radiomic markers that are significantly associated with clinical outcomes.

We propose pathway integrative analysis in Chapter 3. This formulation allows the investigation of how genetic and epigenetic drivers that significantly influence specific pathway activity. We incorporate our analysis into colorectal cancer type

where it has four clear consensus molecular subtypes (CMS) [33]. We target to dig into the pathway driver analysis searching for genomic upstream switches for the key pathways that significantly differentiate CMS groups. In Chapter 4, we further extended the pathway analysis from single cancer type into multiple cancer types where we detect pathway drivers for each cancer type in the group borrowing strength of their potential similarities in the subset of the significant drivers with Bayesian hierarchical modeling framework. We apply our methodology into different cancer groups as in case study, such as Pan-kidney group, Pan-Gyno group [11], Pan-GI group and Pan-Squamous group [18]. Finally, Chapter 5 contains conclusion and future direction with an overview of the novelty, advantages and possible improvement provided by the methodologies presented throughout the dissertation.

Chapter 2

Radio-iBAG: Radiomics-based Integrative Bayesian Analysis of Multiplatform Genomic Data

2.1 Introduction

In oncology, it is of critical importance to investigate both inter- and intra-tumor heterogeneity through an in-depth understanding of the complex interplay between genotypes and phenotypes, towards developing rational anti-cancer therapeutic strategies [26]. The increased availability of complementary and matched molecular and imaging data allows for a thorough examination of tumor heterogeneity at multiple levels [85], [45], [34]. Investigations at the molecular level have been tremendously improved by the development of many genomic profiling technologies, including microarrays, next-generation sequencing, methylation arrays, and proteomic analyses. The Cancer Genome Atlas (TCGA) project, aiming to provide more comprehensive information of human cancer genomes by creating an "atlas" of high-throughput multiple genomic profiles across multiple cancers, was launched

in 2005 as a publicly funded project [108]. The growing availability of such data has motivated the development of integrative analytical models that incorporate various genomic platforms to detect complex patterns of tumor heterogeneity that have predictive and prognostic ability [112].

While genomic data provide information on the molecular characterization of a disease, imaging modalities such as X-ray radiography, magnetic resonance imaging (MRI), computed tomography, and positron emission tomography provide visual and broad resources for the acquisition of high-quality images and provide complementary quantitative information about the structural aspects of a disease. In the context of cancer, these imaging modalities provide a quantitative basis for detailed assessment of various features of the tumor that are associated with the development and progression of cancer. *Radiomics* is an emerging field with a goal of providing predictive or prognostic information by revealing quantitative mechanistic associations between radiologic images and clinical outcomes [23], [2], [27], [64]. Radiomics, in general, involves the extraction and mining of various types of quantitative imaging features that are processed from high-throughput images obtained via different imaging modalities. These imaging features describe different morphological characteristics of a tumor, e.g., tumor shape features such as round or spiculated, total volume or surface area, intensity histogram features that describe the contrast intensity level, and textural features such as energy and entropy that evaluate tumor spatial heterogeneity. In particular, "texture analysis", which applies different statistical models and mathematical transforming methods to further evaluate a tumor's intra-lesional heterogeneity, has become an active ongoing area of research [20]. In the context of glioblastoma multiforme (GBM), several studies have shown that the textural features from perfusion parametric maps provide useful information for predicting patients' survival times [64] and the features extracted from a gray-level

co-occurrence matrix (GLCM) [38], [20] are effective in discriminating tumor volumetric phenotypes [21].

Radiomic and genomic features capture complementary characteristics of the underlying tumor, with radiomics capturing visual phenotypic information in the tumor and genomics capturing its underlying molecular biology. Thus, it is of interest to assess the interrelationships of these two types of features, a task termed *radiogenomics*, and then collectively assess how these inter-related features correlate with clinically relevant endpoints (e.g., survival, progression). From an analytical standpoint, radiogenomic analysis faces several key challenges. First, incorporating complex biological interactive mechanisms, both within and between multiple genomic platforms at the genomic (DNA), transcriptomic (mRNA) and epigenomic (methylation) levels, is understudied in the radiogenomic framework. Second, the high-dimensional nature of both the quantitative features of images and genomic markers necessitates proper dimension reduction techniques and feature selection methods. Third, the analysis becomes more complicated when we wish to link clinical outcomes with genomic and radiomic outcomes in addition to modeling associations between the radiomic and genomic measurements to provide potentially biologically and clinically translatable results.

Multiple studies have addressed these challenges to various degrees. Taking advantage of multi-platform genomic data resources, additive models have been developed that treat the features from different platforms in the same models, although not explicitly modeling their interrelationships [24], [59]. Wang et al., 2013 [112] proposed an integrative Bayesian analysis framework to integrate multi-platform genomic data using hierarchical models that capture the natural mechanistic relationships among the various molecular resolution levels. Jennings et al., 2012 [48]

generalized the method to integrate various types of genomic platforms with a single clinical outcome. These methods effectively capture the biological interaction within different molecular processes, but do not consider high dimensionality in the outcomes. Olivares et al., 2013 [87] extended the above model with multivariate correlated imaging outcomes. This approach models image markers in separate linear models after applying a de-correlating procedure, but does not consider patient-specific clinical outcomes. Stingo et al., 2013 [102] developed an integrative Bayesian modeling approach for imaging-genetics that incorporates the binary disease status as a clinical response, and developed a hierarchical mixture model that can select discriminatory imaging regions of interest and their relevant single-nucleotide polymorphisms simultaneously. Similarly, Batmanghelich et al., 2013 [10] developed a joint probabilistic model of imaging and genetic features associated with disease measures, to provide insights into how imaging biomarkers can serve as intermediate phenotypes when detecting genetic and diagnostic associations. However, these approaches only consider individual platforms and thus do not consider the interrelationships among the various molecular resolution levels in their analytical frameworks.

In this Chapter, we introduce Radio-iBAG: Radiomics-based integrative Bayesian analysis of multiplatform genomic data, an integrative multi-scale Bayesian framework to perform radiogenomic analyses. Our goal is three-fold: first, to detect explicit associations among different genomic platforms at the different molecular levels; second, to treat the radiomic-based biomarkers as an intermediate phenotype (i.e., endo-phenotype), evaluate the molecular underpinnings regulating these biomarkers and finally, evaluate the eventual associations with relevant patient-level clinical outcomes (e.g., survival times). To accomplish these tasks, we construct a multi-level regression-based modeling strategy: a first stage "*genomic model*" detects the complex biological mechanistic relationships among different genomic platforms,

a second stage "*radiogenomic model*" subsequently discovers the underlying associations between gene-platform combinations and radiomic biomarkers. To assess clinical relevance, a third level model "*radiogenomic clinical model*" uncovers the associations between clinical outcomes and genomically-driven radiomic markers.

To address the high dimensionality in both the genomic and radiomic datasets, we utilize Bayesian shrinkage-based priors to achieve sparsity and regularization in the high-dimensional covariate space at various hierarchical levels. Specifically, we employ scale-mixture of normal representations, that allow adaptive shrinkage and borrowing strength within and across the different hierarchical levels. Our methodology is motivated by and applied to a GBM case study, wherein we discover multiple radiomic feature groups significantly associated with patients' survival times along with their mechanism of action through multi-platform genomics.

In Section 2.2, we introduce our modeling scheme, major components, modeling methods and biomarker detection for each modeling stage. In Section 2.3, we illustrate our proposed model on the GBM case study with detailed description of the radiomic features and genomic profile datasets, modeling results and biological interpretations. In Section 2.4, we draw some conclusions and discuss some future extensions and advancements.

2.2 Method: Radio-iBAG Model

2.2.1 Modeling stages

As mentioned above, our core construction of the Radio-iBAG model framework consists of a multi-stage Bayesian hierarchical model. In the *genomic model*, we model the complex biological mechanistic relationship among genomic data from different

platforms capturing information at various molecular resolution levels (e.g., gene expression, copy number and methylation). Subsequently, we carry the information garnered from the *genomic model* into the second stage, the *radiogenomic model*, to parse out the imaging-genomic correlations, which are then included as predictors in the third stage, the *radiogenomic clinical model*. This procedure delineates the image features that directly affect clinical outcomes, as well as those that appear to be modulated by combinations of genomic factors. This construction allows us to discover strong relationships between imaging and genomics data, among the genomic platforms, and identify which appear to be associated with clinical outcome.

Figure 2.1 illustrates the general multi-stage modeling scheme. In the first stage, multiplatform genomic data sets are expressed as data matrices: X_{mRNA}, X_{CN}, X_{miRNA} or X_{Methy}, each with rows as samples and columns as gene-level summaries of the respective platforms. In stage II, we consider radiomic features (RFs) that are preprocessed and extracted from imaging data sets, forming a data frame $\boldsymbol{\mathcal{I}}$ with columns as different features and rows as samples. In the final stage, we incorporate into the model the clinical outcome, denoted as $\boldsymbol{Y}$, which is a vector with the number of elements as the sample size. The construction of each modeling stage is explained in detail in the ensuing sections.

A. *Genomic Model*

Our genomic model involves the integrative modeling of multiplatform genomic data sets. Modern genomics data is comprised of multiple platforms that contain measurements at various molecular resolution levels, from DNA to mRNA to proteins, and including epigenetic levels including alterations like methylation and microRNA

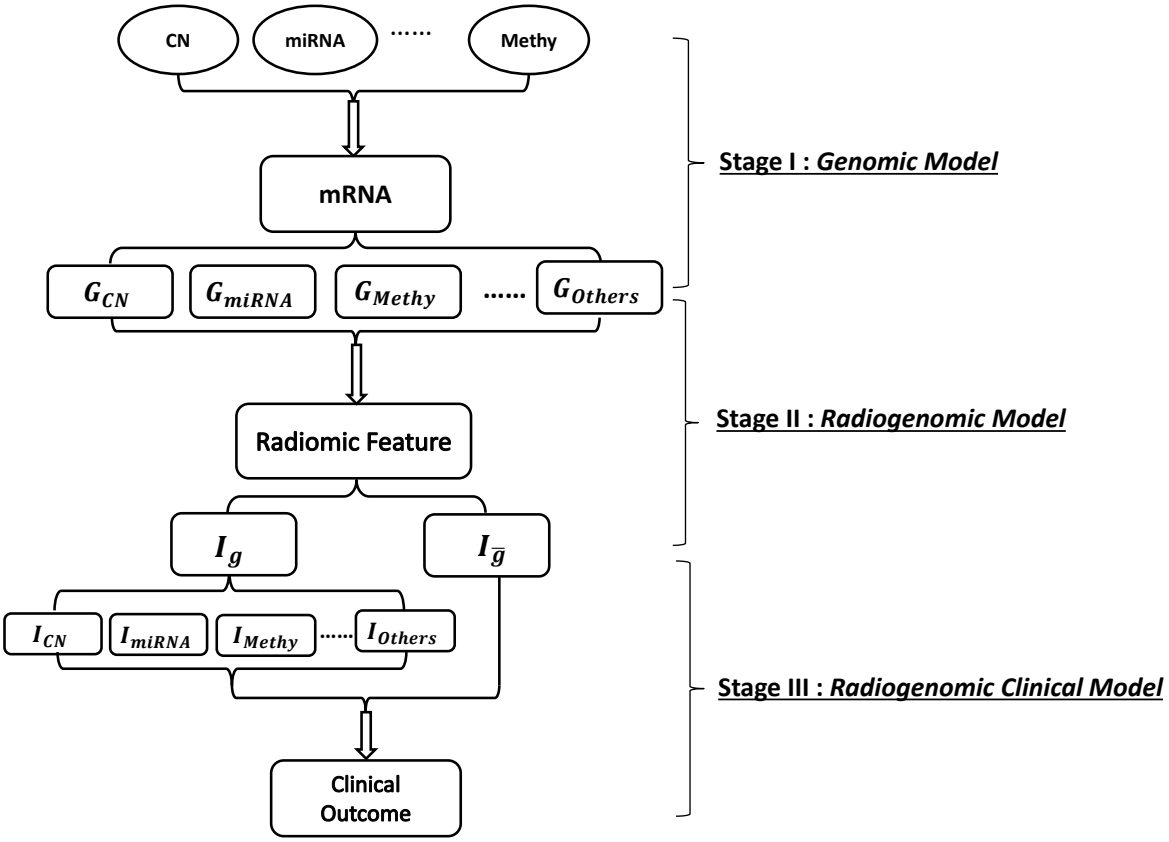


Figure 2.1: Schematic representation of the multi-stage modeling process. In stage I, for each gene, model the relationship between mRNA and different upstream genomic platforms and partition mRNA expression into multiple parts explained by different genomic platforms, CN: copy number alteration, miRNA: microRNA, Methy: methylation, Others: gene expression that is explained by other factors; In stage II, for each radiomic marker, apply Bayesian hierarchical model and partition the radiomic marker into different parts modulated by multiple mRNA factors that are explained by various gene-platform combinations and regard the residual as a non-gene-driven part denoted as $I_{\bar{g}}$; In stage III, apply Bayesian hierarchical model to investigate the relationship between segmented radiomic factors with clinical outcome.

(miRNA) that affect mRNA expression. These platforms capture complementary information at the different molecular resolution levels, and together provide a more complete picture of the underlying biology than any one platform. In this thesis Chapter, we consider three genomic platforms: mRNA, DNA copy number (CN) and miRNA, but the general models we introduce can incorporate any other platforms

capturing upstream genetic and epigenetic information, as well. Also, for a specific gene, we only take the genomic platforms mapped with this gene into our model, we do not consider modeling coexpression or coregulation of the neighboring genes or potential transcriptional regulators. Suppose $N_{\mathcal{G}}$=number of patients with genomic information, J=total number of genomic platforms, and $P_{\mathcal{G}}$=number of target genes. For our particular case, using copy number alteration and miRNA as our upstream platforms, the gene expression level can be modeled and expressed as

$$X_{mRNA_g} = \underbrace{f_1(X_{miRNA_g}) + f_2(X_{CN_g})}_{\text{upstream platform driven}} + \underbrace{\boldsymbol{O_g}}_{\text{explained by other factors}} \tag{2.1}$$

where each $f_j(\cdot)$ is a smooth nonparametric function of the corresponding predictor modeled by a penalized spline formulation that allows us to capture flexible non-linear relationships. We assessed the nonlinearity of gene-level fits and show that GAM provides better fit that GLM for most genes (see Appendix A.3). Other types of splines or alternative nonparametric models could also be used. Our analysis in this stage matches the first stage of the iBAG model [112], whereby the gene expression of a given gene is modeled as explained by upstream factors, with the effects of upstream factors modeled nonparametrically as in [50] via a generalized additive model (GAM) [40]. In principle, the model can include any number of upstream (to mRNA) platform types, including methylation, copy number, loss of heterozygosity, methylation, miRNA, and transcription factors, as long as matched data are available.

The terms in the model are described and interpreted as follows:

- X_{mRNA_g} is the expression of gene g with dimension $N_{\mathcal{G}} \times 1$, $g = 1, 2, ..., P_{\mathcal{G}}$
- X_{miRNA_g} is an aggregated miRNA expression value that integrates information across miRNAs that have been documented to regulate the expression of gene g. For a given gene, there exist multiple miRNAs that interact with this gene,

and here we construct gene-level summaries of these miRNAs that condense their activity into a lower dimension using principal components, as described in detail in Section 2.2.3. The gene-level summaries X_{miRNA_g} have dimension $N_{\mathcal{G}} \times M_{miRNA_g}$, where M_{miRNA_g} denotes the number of gene-level summary vectors for the g^{th} gene.

- X_{CN_g} are gene-level summaries of the CN alteration for the g^{th} gene with dimension $N_{\mathcal{G}} \times M_{CN_g}$. Similarly, as there are multiple CN alteration values from different markers within the same gene, M_{CN_g} denotes the number of gene-level summary vectors.

- O_g represents the "other" part of gene expression that is not captured by the modeled upstream factors, but instead attributed to other upstream factors not in the model, and is of dimension $N_{\mathcal{G}} \times 1$.

This model is fit separately for each gene, and effectively partitions the information contained in the mRNA measurements into an additive set of components, with each component capturing the part of mRNA expression explained by a particular upstream platform. We call these parts different genomic platform components. For gene g, the components can be estimated based on the following formula: $G_{miR_g} = \hat{f}_1(X_{miRNA_g})$, $G_{CN_g} = \hat{f}_2(X_{CN_g})$ and $G_{O_g} = X_{mRNA_g} - \hat{f}_1(X_{miRNA_g}) - \hat{f}_2(X_{CN_g})$. Repeating the same procedure for all the genes, we combine the components grouped by platform, forming different genomic platform combinations: $\mathbf{G_{miR}} = \{G_{miR_1}, G_{miR_2}, ..., G_{miR_{P_{\mathcal{G}}}}\}$, $\mathbf{G_{CN}} = \{G_{CN_1}, G_{CN_2}, ..., G_{CN_{P_{\mathcal{G}}}}\}$ and $\mathbf{G_O} = \{G_{O_1}, G_{O_2}, ..., G_{O_{P_{\mathcal{G}}}}\}$. These combinations represent the gene expression level attributed to miRNA, CN and other factors, respectively, for all $P_{\mathcal{G}}$ target genes of interest.

At times, not all samples with genomic data have radiomic data, as in our GBM

example. In that case, we denote $N_{\mathcal{GI}}$ ($N_{\mathcal{GI}} \subseteq N_{\mathcal{G}}$) as the sample size of their intersection. We carry forward the corresponding subset of the estimated gene platform combinations $\mathbf{G_{miR}}$, $\mathbf{G_{CN}}$, $\mathbf{G_O}$, each with dimension $N_{\mathcal{GI}} \times P_{\mathcal{G}}$, as predictors into the second-stage *radiogenomic model*.

B. *Radiogenomic Model*

The goal of the second stage *radiogenomic model* is to find gene-platform combinations that appear to be associated with radiomic markers, and to partition the radiomic markers into the parts modulated by different gene effects carried from the *genomic model* and those that are not modulated by the modeled genomic factors. The model can be written as

$$\begin{aligned} \boldsymbol{\mathcal{I}} &= \boldsymbol{\mathcal{I}_g} + \boldsymbol{\mathcal{I}_{\bar{g}}} \\ &= \underbrace{\boldsymbol{G_{miR}}\boldsymbol{B_{miR}} + \mathbf{G_{CN}}\boldsymbol{B_{CN}} + \mathbf{G_O}\boldsymbol{B_O}}_{\text{Genomically driven}} + \underbrace{\boldsymbol{\mathcal{I}_{\bar{g}}}}_{\text{Non-genomically driven}} \end{aligned} \tag{2.2}$$

The terms in the model can be expressed and interpreted as follows:

- $\boldsymbol{\mathcal{I}}$ denotes a $N_{\mathcal{GI}} \times K$ matrix in which K is the number of general RFs (individual radiomic features or Radiomic-meta-Features (RmFs) that we constructed from high dimensional RFs that are highly correlated, described in detail in section 3.2).

- $\boldsymbol{B_{miR}}$ is of dimension $P_{\mathcal{G}} \times K$, with columns as the vectors of the expression effects for corresponding radiomic markers through miRNA;

- $\boldsymbol{B_{CN}}$ is of dimension $P_{\mathcal{G}} \times K$, with columns as the vectors of the expression effects for corresponding radiomic markers through CN;

- $\boldsymbol{B_O}$ is of dimension $P_{\mathcal{G}} \times K$, with columns as the vectors of the expression effects for the corresponding radiomic markers through "other" genomic mechanistic factors;

- $\mathbf{G_{miR}}$, $\mathbf{G_{CN}}$, $\mathbf{G_O}$ are the estimated gene expression components described in part A.

Associations are detected by examining the coefficients' posterior probabilities based on Markov chain Monte Carlo (MCMC) samples, and estimates given by posterior means. To achieve the segmentation of the radiomic features, we can estimate each component by $\hat{\boldsymbol{\mathcal{I}}}_{\boldsymbol{CN}} = \mathbf{G_{CN}}\hat{\boldsymbol{B}}_{\boldsymbol{CN}}$, $\hat{\boldsymbol{\mathcal{I}}}_{\boldsymbol{miR}} = \mathbf{G_{miR}}\hat{\boldsymbol{B}}_{\boldsymbol{miR}}$, $\hat{\boldsymbol{\mathcal{I}}}_{\boldsymbol{O}} = \mathbf{G_O}\hat{\boldsymbol{B}}_{\boldsymbol{O}}$. The final non-gene-driven part can be estimated by $\hat{\boldsymbol{\mathcal{I}}}_{\bar{\boldsymbol{g}}} = \boldsymbol{\mathcal{I}} - \mathbf{G_{CN}}\hat{\boldsymbol{B}}_{\boldsymbol{CN}} - \mathbf{G_{miR}}\hat{\boldsymbol{B}}_{\boldsymbol{miR}} - \mathbf{G_O}\hat{\boldsymbol{B}}_{\boldsymbol{O}}$. We then further carry the above four components into the final stage, the *radiogenomic clinical model*.

C. *Radiogenomic Clinical Model*

The third-stage model relates the various radiogenomic marker combinations from the second stage model to a clinical outcome (e.g., survival time in our context). The model can be expressed as

$$\boldsymbol{Y} = \boldsymbol{\mathcal{I}}_{\boldsymbol{CN}}\boldsymbol{\alpha_1} + \boldsymbol{\mathcal{I}}_{\boldsymbol{miR}}\boldsymbol{\alpha_2} + \boldsymbol{\mathcal{I}}_{\boldsymbol{O}}\boldsymbol{\alpha_3} + \boldsymbol{\mathcal{I}}_{\bar{\boldsymbol{g}}}\boldsymbol{\alpha_4} + \boldsymbol{\epsilon},$$

where Y is the clinical outcome with dimension $N_{\mathcal{GIC}} \times 1$ and $N_{\mathcal{GIC}}$ ($N_{\mathcal{GIC}} \subseteq N_{\mathcal{GI}} \subseteq N_{\mathcal{G}}$) is the sample size of the intersection of the genetic, image and clinical data sets. $\boldsymbol{\mathcal{I}}_{\boldsymbol{CN}}$ is the CN modulated radiomic marker component matrix. Similarly, $\boldsymbol{\mathcal{I}}_{\boldsymbol{miR}}$ denotes the microRNA modulated part; $\boldsymbol{\mathcal{I}}_{\boldsymbol{O}}$ is the part of radiomic features explained by a genomic factor but modulated by something other than CN or miRNA; and $\boldsymbol{\mathcal{I}}_{\bar{\boldsymbol{g}}}$ denotes the part of the radiomic feature not regulated by genes in the model. All

four radiomic marker components have the dimension $N_{\mathcal{GIC}} \times K$. $\boldsymbol{\alpha_1}$, $\boldsymbol{\alpha_2}$, $\boldsymbol{\alpha_3}$, $\boldsymbol{\alpha_4}$ denote the corresponding image marker combination effects. $\boldsymbol{\epsilon}$ is the error term for modeling the clinical outcome. In our GBM application, where the clinical outcome is survival time, we use an accelerated failure time (AFT) model, with Y as the log-transformed survival time [113]. However, for the general analytical process, our outcome $\boldsymbol{Y}$ can involve any clinical measurements with suitable regression model determined by the type of outcome (e.g., logistic models for binary outcomes or Cox proportional hazards models in the presence of censored outcomes.)

Our goal in this final stage is to identify radiomic markers associated with clinical outcome, either modulated by genomic factors or not. We identify these factors by estimation and Bayesian posterior inference of $\boldsymbol{\alpha} = \{\boldsymbol{\alpha_1}, \boldsymbol{\alpha_2}, \boldsymbol{\alpha_3}, \boldsymbol{\alpha_4}\}$, and then can characterize these effects in more detail by tracing information back through the earlier stage models. For example, if a particular radiomic feature is related to clinical outcome through a genomic effect, we can examine the corresponding second stage model to identify which genes are driving such effects, and then the first stage model for those genes to find which upstream platforms most strongly modulate the expression of those genes. In this way, the radio-iBAG model can not only detect clinically relevant radiomic features, but provide a thorough summary of the radiomic-genomic and multi-platform genomic interrelationships that appear to modulate these factors.

2.2.2 Radio-iBAG Model Estimation

Our second- and third-stage models involve multiple genes and/or RFs, so it is necessary to introduce sparsity into the regression models to regularize the fitting and to obtain a relatively smaller and more interpretable set of radiogenomic factors that appear to be related to the clinical outcome. This can be done using penalized

likelihood or other regularization techniques, but here we use a Bayesian approach and induce sparsity through the prior distributions on the regression parameters.

Some commonly used sparsity priors involve a discrete-mixture prior consisting of a point mass at zero for noise and a continuous density distribution for signals, for example a normal distribution and a point mass at zero [75]. Other types of sparsity priors do not have a zero component, but instead are absolutely continuous distributions that accomplish sparsity via nonlinear shrinkage, which can often be accomplished using a normal scale mixture prior distribution. Examples include a normal-exponential (Bayesian lasso) [88], Horseshoe [19], generalized double pareto [6], Dirichlet Laplace [13], and Normal Gamma [32]. Considering incorporating the multi-scale property of the datasets by allowing common platform share the same hyperparameters with proper biological interpretation, we seek for prior settings that yield to more direct incorporation. Thus, we mainly consider Bayesian Lasso and Normal Gamma prior settings. While the Bayesian lasso, which is a Bayesian analog to the commonly-used lasso [107], is commonly used, it has limited flexibility given it is determined by a single hyperparameter that regulates both sparsity and the tails. We instead use the normal-gamma (NG) prior [32], which contains a second hyperparameter, and thus can better handle sparsity as well as flexibility to manage the tails and yield to more accurate coefficient estimates, as described and illustrated via multiple simulation settings in [32]. We apply this prior in both stage II *radiogenomic* and stage III *radiogenomic clinical* models. Further, we allow the sparsity hyperparameters to be indexed by platform, which enables borrowing of strength across genes in determining the desired sparsity and tail levels on a platform specific basis.

To estimate the coefficient vector, for the k^{th} RF, we assign the NG prior dis-

tribution to $\boldsymbol{\beta_k} = \{\boldsymbol{\beta^k_{miR}}, \boldsymbol{\beta^k_{CN}}, \boldsymbol{\beta^k_O}\}$, each part of the coefficient vector being assigned with a particular set of the hyperparameters. In this way, we allow priors settings that incoporate multi-scale datasets. More specifically, suppose our genomic platform combination predictors can be expressed as $X = \{\mathbf{G_{miR}}, \mathbf{G_{CN}}, \mathbf{G_O}\}$, then the linear regression model and its hierarchical prior setting can be expressed as

$$\boldsymbol{\mathcal{I}_k} = X\boldsymbol{\beta_k} + \boldsymbol{\mathcal{I}_{k\bar{g}}}$$

$$\boldsymbol{\mathcal{I}_k} \sim Normal(X\boldsymbol{\beta_k}, \sigma_k^2 \boldsymbol{I_{N_{\mathcal{GI}}}})$$

$$\boldsymbol{\beta_k} \sim Normal(\mathbf{0}_{\tilde{\boldsymbol{P}}}, D_\psi)$$

$$D_\psi = diag(\psi_{1,1}, \psi_{1,2}, ..., \psi_{1,P_1}, \psi_{2,1}, \psi_{2,2}, ..., \psi_{2,P_2}, ..., \psi_{J,1}, \psi_{J,2}, ..., \psi_{J,P_J}),$$

where $\tilde{P} = P_1 + P_2 + ... + P_J$ is the total number of predictors (dimension of X), J denotes the total number of platform types ($j = 1, 2, 3, ..., J$, here our $J = 3$), and P_j denotes the total number of genomic features (each sub-indexed as g) for the j^{th} genomic platform type. Our estimation of the scale parameters and the main coefficients ($\boldsymbol{\beta_k}$) is processed by applying the NG prior $\psi_{j,g} \sim Gamma(\lambda_j, 1/(2\gamma_j^2))$ for the j^{th} platform. Also, the hyper-prior $\lambda_j \sim exp(c)$ and $\gamma_j^{-2} \sim Gamma(\tilde{a}, \tilde{b}/(2\lambda_j))$ are assigned to induce greater flexibility and completeness in shrinkage estimation. To complete our prior specification, we assume a conjugate $InverseGamma(a, b)$ prior on σ_k^2. Here, we let each genomic platform combination (platform type) share the same set of hyperparameters (λ_j, γ_j^2), thus maintaining the grouped structure at the shrinkage level. For implementation, we utilize Markov Chain Monte Carlo (MCMC) based Bayesian sampling techniques such as Gibbs sampling and Metropolis-Hastings. The posterior means calculated from MCMC samples are used to obtain the parameter estimations, and the corresponding posterior probabilities are used to conduct signal detection. The details for the posterior distribution and

MCMC sampling are shown in Appendix A.1.

For the *radiogenomic clinical model*, we utilize similar NG prior distributions, the only difference being that our group structure is determined by the RF combinations. We assign the same hyperparameters for the partitioned RFs that belong to the same combination/group. Suppose our predictor set estimated from stage II can be expressed as $\boldsymbol{\mathcal{I}} = \{\boldsymbol{\mathcal{I}_{CN}}, \boldsymbol{\mathcal{I}_{miR}}, \boldsymbol{\mathcal{I}_{O}}, \boldsymbol{\mathcal{I}_{\bar{g}}}\}$, and the effect parameter $\boldsymbol{\alpha} = \{\boldsymbol{\alpha_1}, \boldsymbol{\alpha_2}, \boldsymbol{\alpha_3}, \boldsymbol{\alpha_4}\}$, then the model and prior construction can be expressed as

$$\boldsymbol{Y} = \boldsymbol{\mathcal{I}\alpha} + \boldsymbol{\epsilon}$$

$$\boldsymbol{Y} \sim Normal(\boldsymbol{\mathcal{I}\alpha}, \sigma^2 \boldsymbol{I_{N_{\mathcal{GIC}}}})$$

$$\boldsymbol{\alpha} \sim Normal(\mathbf{0}, D_\psi)$$

$$D_\psi = diag(\psi_{1,1}, \psi_{1,2}, ..., \psi_{1,K}, \psi_{2,1}, \psi_{2,2}, ..., \psi_{2,K}, ..., \psi_{J,1}, \psi_{J,2}, ..., \psi_{J,K}),$$

where J denotes the total number of different RF combination types ($j = 1, 2, 3, ..., J$, our $J = 4$), k denotes the RF index ($k = 1, 2, 3, ..., K$). Further, we assign our prior and hyper-prior distributions as $\psi_{j,k} \sim Gamma(\lambda_j, 1/(2\gamma_j^2))$, $\sigma^2 \sim InverseGamma(u_1, u_2)$, $\lambda_j \sim exp(d)$, and $1/(2\gamma_j^2) \sim Gamma(\tilde{e}, \tilde{f}/(2\lambda_j))$. Note that for censored sample i, we sample $\boldsymbol{Y_i}$ from complete conditional distribution which is normal distribution with left truncation at t_i that represents the follow-up time. Finally, RF combination selection is based on the posterior probability of the MCMC samples. Details about the posterior distribution and sampling methods are provided in Appendix A.1.

2.2.3 Radiomic and Genomic Marker Selection

For marker/feature selection we propose a thresholding procedure for the various regression models. Specifically for the *radiogenomic clinical model*, we choose a thresholding criteria considering both the effective size and clinical interpretability. For example, in the GBM case study, we apply the AFT model with the log-transformed survival time as the clinical outcome. In our analysis, considering that the survival times are measured in months, which is comparatively small, we choose to apply $\log_2$-based transformation, which leads to better interpretability and a simpler calculation. Based on this setting, the region for detecting the coefficients of the image markers becomes $\alpha_{jk} \in (-\infty, \delta_-^*) \cup (\delta_+^*, \infty)$, where we denote δ_-^* as $log_2(1-\delta_2)$ and δ_+^* as $log_2(1+\delta_2)$, particularly, α_{jk} is the coefficient of the k^{th} radiomic marker modulated by the j^{th} genomic platform ($j = 1, 2, ..., J$, $k = 1, 2, ..., K$). Moreover, δ_2 is determined to achieve the proper effect size and is interpreted as the percentage change in survival time, e.g., for the GBM data analysis, we choose $\delta_2 = 0.05$, which corresponds to 5% change in survival time. More specifically, we denote $P_+(\mathcal{I}_{jk}) = \sum_{t=S+1}^{t=T} \boldsymbol{I}(\alpha_{jk}^{(t)} > \delta_+^*)/(T-S)$ and $P_-(\mathcal{I}_{jk}) = \sum_{t=S+1}^{t=T} \boldsymbol{I}(\alpha_{jk}^{(t)} < \delta_-^*)/(T-S)$ where t denotes the t^{th} MCMC iteration, S denotes the burn-in sample size and T represents the total number of MCMC iterations. We flag $\mathcal{I}_{jk}$ to be positively significant if $P_+(\mathcal{I}_{jk}) > 0.5$ or negatively significant if $P_-(\mathcal{I}_{jk}) > 0.5$ [9].

Analogously, for the *radiogenomic model*, considering δ-fold or larger variation in the response for a unit change in a particular predictor is defined as a standard in the significance detection, which corresponds to $\beta_{jg} \in (-\infty, -\delta) \cup (\delta, \infty)$ and β_{jg} is the coefficient of the j^{th} platform of the g^{th} gene in the analysis. Once a proper threshold δ_1 is determined, the posterior probability is defined as $P(x_{jg}) = \sum_{t=S+1}^{t=T} \boldsymbol{I}(|\beta_{jg}^{(t)}| > \delta_1)/(T-S)$, where S is the burn-in sample size and T is

the total number of MCMC iterations. Feature x_{jg} in the gene-platform combinations is highlighted to be 'significant' if $P(x_{jg}) > 0.5$.

Radio-iBAG modeling algorithm provides a concise summary of Radio-iBAG model implementation and genomic/radiomic marker selection.

2.3 Radiogenomic Mapping of Glioblastoma Multiforme

Glioblastoma Multiforme (GBM) is an aggressive and malignant form of primary brain cancer. It is the highest grade glial tumor, with a median survival time of 14.6 months following standard treatment options and typically 3 months without treatment [103]. Although different treatment approaches that include radiation, surgery and chemotherapy have been developed and applied in clinical practice, the overall mortality rate still remains high, mainly due to the tumor's resistance to treatment [14] and the complexity of its primary biological mechanism.

Currently, at the molecular level, TCGA provides data sets with multiple genomic platforms, including methylation, CN alteration, and gene expression. Studies based on TCGA platform have identified distinct molecular subclasses of GBM, resembling stages in neurogenesis that are relevant to prognosis [110]. Also, with the availability of standardized medical image annotations from The Cancer Imaging Archive (TCIA), multiple studies currently focus on the detection of radiomic imaging variables associated with clinical outcomes [21], [55]. Relevant studies have shown that quantitative imaging features extracted from different modalities provide strong prognostic information [85], [64].

Radio-iBAG modeling algorithm

Stage I: *Genomic Model*

for each gene g **do**

$X_{mRNA_g} = f_1(X_{miRNA_g}) + f_2(X_{CN_g}) + O_g$

Estimate $G_{miR_g} = \hat{f}_1(X_{miRNA_g})$, $G_{CN_g} = \hat{f}_1(X_{CN_g})$

and $G_{O_g} = X_{mRNA_g} - \hat{f}_1(X_{miRNA_g}) - \hat{f}_2(X_{CN_g})$

end for

aggregate: $\boldsymbol{G_{miR}} = \{G_{miR_1}, G_{miR_2}, ..., G_{miR_{PG}}\}$; similarly for $\boldsymbol{G_{CN}}$ and $\boldsymbol{G_O}$

Stage II: *Radiogenomic Model*

for each RF **k do**

$\boldsymbol{\mathcal{I}_k} = \boldsymbol{G_{miR}}\boldsymbol{\beta^k_{miR}} + \mathbf{G_{CN}}\boldsymbol{\beta^k_{CN}} + \mathbf{G_O}\boldsymbol{\beta^k_O} + \boldsymbol{\mathcal{I}_{k\bar{g}}} = X\boldsymbol{\beta_k} + \boldsymbol{\mathcal{I}_{k\bar{g}}}$

MCMC sampling of β_{jg} (j: platform; g: gene index) for T iterations.

Calculate posterior probability with burn-in sample size S.

if $P(x_{jg}) = \sum_{t=S+1}^{t=T} \boldsymbol{I}(|\beta_{jg}^{(t)}| > \delta_1)/(T-S) > 0.5$ **then**

x_{jg} (g: gene index; j: platform index) is flagged as important

end if

Estimate β_{jg} by posterior mean: $\hat{\beta}_{jg} = \frac{1}{T-S}\sum_{t=S+1}^{t=T} \beta_{jg}^{(t)}$

segment $\hat{\boldsymbol{\beta}}_{\boldsymbol{k}} = \{\hat{\beta}_{11}, \hat{\beta}_{12}, ..., \hat{\beta}_{JP_J}\}^T = \{\hat{\boldsymbol{\beta}}^{\boldsymbol{k}}_{\boldsymbol{miR}}, \hat{\boldsymbol{\beta}}^{\boldsymbol{k}}_{\boldsymbol{CN}}, \hat{\boldsymbol{\beta}}^{\boldsymbol{k}}_{\boldsymbol{O}}\}^T$

Thus $\hat{\boldsymbol{\mathcal{I}}}_{\boldsymbol{kg}} = \boldsymbol{G_{miR}}\hat{\boldsymbol{\beta}}^{\boldsymbol{k}}_{\boldsymbol{miR}} + \mathbf{G_{CN}}\hat{\boldsymbol{\beta}}^{\boldsymbol{k}}_{\boldsymbol{CN}} + \mathbf{G_O}\hat{\boldsymbol{\beta}}^{\boldsymbol{k}}_{\boldsymbol{O}} = \hat{\boldsymbol{\mathcal{I}}}^{\boldsymbol{k}}_{\boldsymbol{miR}} + \hat{\boldsymbol{\mathcal{I}}}^{\boldsymbol{k}}_{\boldsymbol{CN}} + \hat{\boldsymbol{\mathcal{I}}}^{\boldsymbol{k}}_{\boldsymbol{O}}$

and non-gene-driven part $\hat{\boldsymbol{\mathcal{I}}}_{\boldsymbol{k\bar{g}}} = \boldsymbol{\mathcal{I}_k} - \hat{\boldsymbol{\mathcal{I}}}_{\boldsymbol{kg}}$

end for

aggregate: $\boldsymbol{\mathcal{I}_{miR}} = \{\hat{\boldsymbol{\mathcal{I}}}^{\boldsymbol{1}}_{\boldsymbol{miR}}, \hat{\boldsymbol{\mathcal{I}}}^{\boldsymbol{2}}_{\boldsymbol{miR}}, ..., \hat{\boldsymbol{\mathcal{I}}}^{\boldsymbol{K}}_{\boldsymbol{miR}}\}$; $\boldsymbol{\mathcal{I}_{CN}} = \{\hat{\boldsymbol{\mathcal{I}}}^{\boldsymbol{1}}_{\boldsymbol{CN}}, \hat{\boldsymbol{\mathcal{I}}}^{\boldsymbol{2}}_{\boldsymbol{CN}}, ..., \hat{\boldsymbol{\mathcal{I}}}^{\boldsymbol{K}}_{\boldsymbol{CN}}\}$;
$\boldsymbol{\mathcal{I}_O} = \{\hat{\boldsymbol{\mathcal{I}}}^{\boldsymbol{1}}_{\boldsymbol{O}}, \hat{\boldsymbol{\mathcal{I}}}^{\boldsymbol{2}}_{\boldsymbol{O}}, ..., \hat{\boldsymbol{\mathcal{I}}}^{\boldsymbol{K}}_{\boldsymbol{O}}\}$ and $\boldsymbol{\mathcal{I}_{\bar{g}}} = \{\hat{\boldsymbol{\mathcal{I}}}_{\boldsymbol{1\bar{g}}}, \hat{\boldsymbol{\mathcal{I}}}_{\boldsymbol{2\bar{g}}}, ..., \hat{\boldsymbol{\mathcal{I}}}_{\boldsymbol{K\bar{g}}}\}$

Stage III: *Radiogenomic Clinical Model*

predictor matrix $\boldsymbol{\mathcal{I}} = \{\boldsymbol{\mathcal{I}_{CN}}, \boldsymbol{\mathcal{I}_{miR}}, \boldsymbol{\mathcal{I}_O}, \boldsymbol{\mathcal{I}_{\bar{g}}}\}$

coefficient vector $\boldsymbol{\alpha} = \{\boldsymbol{\alpha_1}, \boldsymbol{\alpha_2}, \boldsymbol{\alpha_3}, \boldsymbol{\alpha_4}\}$

modeling: $\boldsymbol{Y} = \boldsymbol{\mathcal{I}\alpha} + \boldsymbol{\epsilon}$

MCMC sampling of α_{jk} (j: RF combination group, k: RF index within each group) for T iterations.

Calculate posterior probability with burn-in sample size S.

if $P(\mathcal{I}_{jk}) = \sum_{t=S+1}^{t=T} I(|\alpha_{jk}^{(t)}| > \delta_2)/(T-S) > 0.5$ **then**

$\mathcal{I}_{jk}$ (j: RF combination group index; k: RF index) is flagged as significant.

The availability of such large-scale data resources (TCIA and TCGA) makes it feasible to perform radiogenomic mapping in GBM to explore the complex associations between molecular features and imaging features for this particular cancer type. In this section, we apply our integrative multi-stage Bayesian hierarchical model with the data from patients with GBM and matched with TCGA and TCIA platforms, to discover radiogenomic associations characterizing these data and identify RmFs and genomic markers associated with GBM prognosis. More details of the genomic and imaging data sets are provided hereafter.

2.3.1 Data Description

2.3.1.1 Radiomic and clinical data description

Among 304 GBM patients with available genomic records, 78 matched patients ($N_{\mathcal{GI}}$=78) also have MRI T1-weighted post contrast images and T2-weighted fluid attenuated inversion recovery (T2-weighted FLAIR) images available from TCIA for texture analysis. Image preprocessing procedures, including steps such as non-uniformity normalization (N3) correction, registration, segmentation, isotropic voxel-reslicing and image filtering, were performed prior to texture feature extraction. For this analysis, we derived textural features from the axial 2D slice that has the largest tumor area [122]. Our textural features were obtained from a two-step process: 1) Image filtering, 2) Haralick features[1]derivation [38] [36] and summary measures calculation. These image pre-processing steps as well as the texture feature calculations are described in detail in the Appendix Section A.2.

For the radiomic data set, we had 972 RFs that could be categorized into 20 groups based on how they were calculated. The group names and corresponding

[1]Features generated using various metrices of the co-occurence matrices are called "Haralick features" after the publication of [38].

descriptions are provided in the Appendix Table A.3. They cover the features of both T2-weighted FLAIR and T1-weighted post-contrast MRI modalities with different type of features: texture features, histogram features and regional features and with two types of ratio based normalization methods.

For clinical outcomes, we utilized overall survival times (in months) as the response in our integrative analysis. For the clinical model, we used data from $N_{\mathcal{GIC}} = N_{\mathcal{GI}} = 78$ GBM patients with matching multi-platform genomic, radiomic, and clinical data, and with 9 patients having censored clinical outcomes. We applied the AFT model using the $\log_2$ transformed survival time $\log_2(T_i)$ as the response, where T_i is the survival time in months after diagnosis for patient i, and imputed the survival time for censored samples simultaneously.

2.3.1.2 Genomic data description

Our gene expression data set is level 3 (summarized per gene), and was downloaded and processed by TCGA Assembler [123] with open-source software and related instructions available in public. The CN data set is level 2 (probe-level) data obtained from TCGA Portal from the HG_CGH_244A platform with normalized records of CN alteration for each probe. The miRNA data set was also acquired from TCGA Portal with 534 miRNA records and 575 samples in total.

In our analysis, we focus on 49 genes that are members of signaling pathways that have previously been detected associated with GBM (RTK/PI3K, P53, and RB pathways [84]) and 304 patients ($N_{\mathcal{G}} = 304$) with records available for mRNA, CN and miRNA. The sample sizes and the specific types of the raw datasets for all genomic platform, radiomic data and clinical data are illustrated via diagram in Appendix Figure A.3 with description in Section A.2.3. The genomic datasets used

in the first stage are all continuous and the descriptions of the raw data structure (for 304 samples) of different genomic platforms are given in below:

- mRNA (304×49) contains gene expression levels for each gene and each patient.
- Copy number (304×491) contains the CN alteration data (columns) for each sample (rows). There exist multiple copy number markers per gene, and the columns of the data set are sorted by gene. Also, one gene, HRAS, does not have CN alteration information, thus, any variance of gene expression contributed by CN changes will be captured by the factor "others" in this analysis; in other words, for gene HRAS, the corresponding column in matrix $\boldsymbol{G_{CN}}$ is set as zero.
- miRNA (304×522) contains miRNA values for each gene (column) and patient (row) based on the miRNA-mRNA interaction membership matrix, with records coming from targetHub [69], which collected miRNA-mRNA interaction records based on 5 external databases, and multiMiR [93] is based on 14 external databases, including validation databases, prediction databases and drug-associated databases. There exist multiple miRNA records corresponding to one gene, and the columns of the miRNA data set are ordered by gene.

We wish to obtain gene-level summaries for each platform based on these raw data sets. Considering that a given gene can contain multiple values from different markers for both miRNA and CN alteration records, and including all these records into the *genomic model* is computationally expensive and inefficient, the gene-level summaries that can be carried into the modeling stage need to be generated. There are different ways to obtain gene-level summaries, e.g., taking the average, selecting the top most correlated records, or extracting the top principal components via PCA. For the analysis of GBM data, CN alteration and miRNA, in each case, we perform PCA on the genomic platform data set mapped to a gene and keep the top principal components with cumulative variance that explain up to 90% of the

total variance. In this way, we regard the remaining records as capturing most of the information of the genomic platform data. Specifically, for gene g, the gene-level summaries for each platform can be expressed as X_{miRNA_g} and X_{CN_g}, which have been denoted in Methods. Our genomic model is conducted based on these three data sets, X_{mRNA_g},X_{miRNA_g} and X_{CN_g}.

As described in Section 2.2.1, our *genomic model* uses the GAM to fit the model and estimate the partitioned mRNA that is modulated by different genomic platforms. To implement the GAM algorithm, we utilized Wood's R package "mgcv" and exploited its option for the automatic smoothness selection for the penalty parameter based on generalized cross-validation [117]. Subsequently, for each gene, we calculated the proportion of the mRNA variance explained by each platform. We assume that if a genomic platform does not explain much variation in mRNA expression, it will not have a significant impact on image features. Thus, for $\mathbf{G_{miR}}$, $\mathbf{G_{CN}}$ and $\mathbf{G_O}$, we filtered out the genomic platform features that explain less than 10% of the total variance of gene expression, leaving the remaining features to be carried forth into the *radiogenomic model*.

2.3.2 Estimation of Radiomic-meta-Features

One of the critical challenges in fitting the radiogenomic model is the high dimensionality and redundancy of the set of radiomic features (RFs). In our GBM case study, the preprocessed RF data set has 972 features, and contains many features within the same type of radiomic class but with different settings, e.g. filtering scales. Thus, there are extensive correlation among many RFs with high magnitudes up to 0.99, as can be seen in the correlation heatmap shown in Appendix Figure A.1. Facing these challenges, we utilize a new radiomic strategy of empirically constructing radiomic meta features (RmFs) comprised by a linear combination a sparse subset of highly

correlated RFs. Each RmF defines a factor capturing one aspect of the fundamental structure in the radiomic features, and together the relatively small number of RmFs retain a vast majority of information contained in the set of 972 RFs. To our knowledge, this strategy has not been applied in the radiomics literature to date, and may be useful in other contexts. We construct the RmFs by applying sparse principal component analysis (sPCA) [125] which incorporates a regularization technique such as the lasso or elastic net to induce sparsity in the principal component loadings. This has the advantage of interpretability over general principal components that do not in general yield sparse loadings, in the case of our GBM application yielding RmFs that are reasonably intuitive and interpretable (see Section 2.3).

This algorithm offers a parsimonious way to obtain more comprehensive representation of radiomic features, which contain the maximum information of the original radiomic data. While not strictly orthogonal like PCs, the SPCs are approximately orthogonal so it is reasonable to model these RmFs as independent imaging features in the second stage radiogenomic model. The sparse loadings for the RmFs for our GBM application are shown in Figure 2.2, and by contrast, the non-sparse loadings for ordinary PCA are shown in the Appendix A.2.2 Figure A.2.

Let $\mathcal{M}$ be an $N_{\mathcal{GI}} \times P$ matrix (typically with $P >> N_{\mathcal{GI}}$) with the rows being the subjects and the columns the P (=972) RFs. The sparse PCA is applied as follows:

- Apply ordinary PCA to $\mathcal{M}$ and record the number of top principal components with the cumulative variance explaining up to $100(1-\alpha)\%$ (eg. 90%) of the total variance. Each PC is regarded as a linear combination of the original features with its loadings can be estimated by regressing the PC on these features. Sparsity in loadings results from adding regularization terms in the regressions.

- The general sPCA algorithm and its numerical computation procedure are described by [125]. In most cases, the number of features is typically much bigger than the sample size; hence, the simplified version of the general sPCA described in the this Chapter should be applied here. The mathematical formulation of sPCA is illustrated in the Appendix A.2.2. To implement the algorithm, we utilized the R package "elasticnet" [124], with K (the number of principal components based on the ordinary PCA) principal components and vectors of λ_j (L1 norm regularization parameter for each loading vector). The parameter λ_j can be chosen by cross validation, or various values can be tried to find one that results in the desired level of sparsity.

Suppose V is our final matrix of loadings with dimensionality $P \times K$, the projected imaging features matrix (PC score matrix) is then $\mathcal{I}_{(N_{\mathcal{GI}} \times K)} = \mathcal{M}_{(N_{\mathcal{GI}} \times P)} V_{(P \times K)}$. We define the vectors of this matrix as RmFs, which contain the majority of the information of the original radiomic data. These features are further regarded as predictors in the analysis of the *radiogenomic clinical model*.

2.3.3 Results Using the Radio-iBAG Model

2.3.3.1 Radiomic-meta-Feature Estimation

We conducted sPCA with the regularization parameter $\lambda = 2.5$ for each principal component, leading to 22 top principal scores that explain 80.7% of the total variance. Through the exploration of the results utilizing different λ values, we finally chose $\lambda = 2.5$ given its balance in the sparsity of the loadings which leads to good interpretation and the cumulative variance that could be attained. We call these 22 principal scores Radiomic-meta-features (RmFs) as discussed in section 3.2. To summarize the RmFs, plots a heatmap of the squared loading proportions within the 20 broad categories of RFs to show which feature types dominate each RmF.

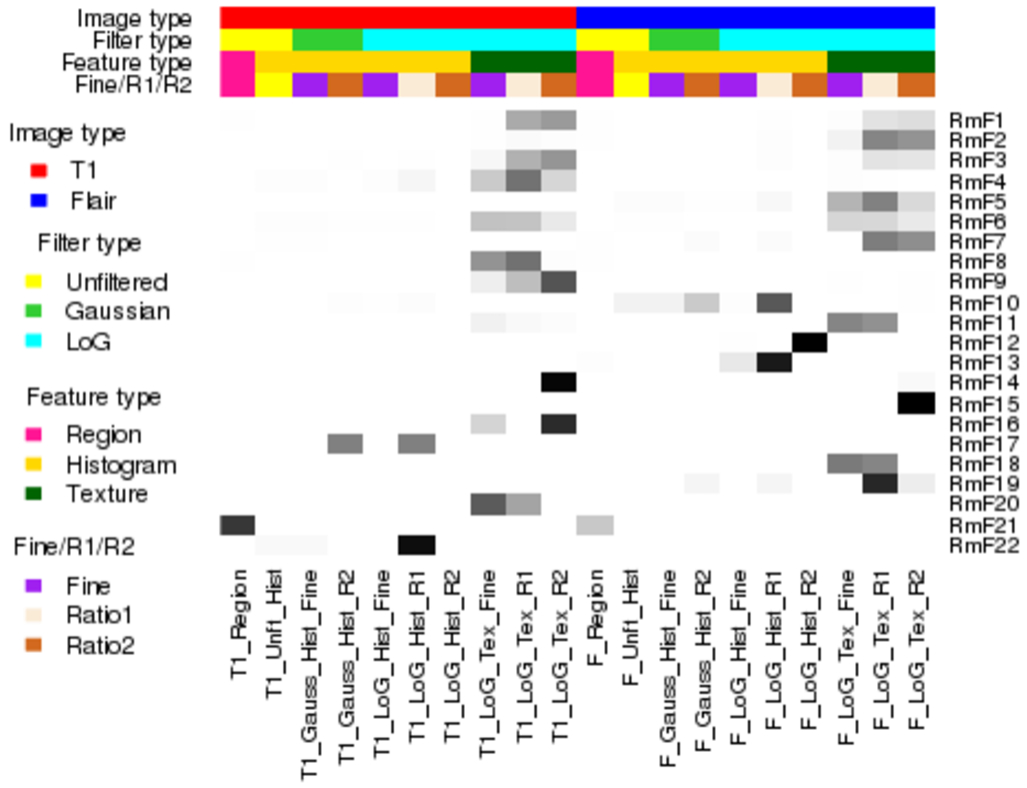


Figure 2.2: Squared loading proportion for each RF group. For each of the 22 radiomic-meta-features (RmFs), the sum of the squared loadings of each group is calculated, divided by the total sum of the squared loadings, which equals exactly 1. The heatmap shows this values in grey level, interpreted the RF group importance for each RmF. The grey level ranging from white to black matches the proportional values ranging from 0 to 1.

This figure reveals that many of the RMFs appear to be interpretable in the sense of summarizing certain aspects of the images, including morphological imaging features that can be directly visualized, eg. unformity, tumor area, mean intensity, etc. To further illustrate their interpretability, we pick out three example RmFs and in Figure 2.3 plot T1-Post Contrast images for the four tumors with highest and lowest values of the corresponding RmF scores, rescaled to [0,1]. RmF 21 has the largest loading values for feature categories indicating tumor area (T1_Region, F_Region). The first column of Figure 2.3 shows that samples with higher values of RmF 21

tend to have larger tumor area. RmF 14 has non-zero loadings inversely proportional to pixel intensity variance measures, and thus can be construed as representative of local pixel heterogeneity. From the second column, it is evident that larger RmF 14 (smaller variance) leads to lower local pixel heterogeneity. The third column of Figure 2.3 shows the sorted RmF 17, whose loadings are dominated by the imaging intensity histogram feature "uniformity", which represents how non-uniform of the overall gray-level pixel intensities. The gray level of the magnified tumor region shows that when RmF value gets larger, the tumor surface gets more non-uniform. These RmFs quantitatively capture these fundamental features of the images.

We use these RmF as quantifications of the radiomic data in our modeling, with the radiomic model fit to Rmf matrix $\boldsymbol{\mathcal{I}}$, which is of dimension 78×22, with RmFs as columns and subjects as rows.

2.3.3.2 Radio-iBAG Modeling Results

Our model shows proper convergence and it is not sensitive to the choice of the hyperparameters based on the model checking results respectively described and shown in Appendix A.4.2. After model fitting, the information about the prognostic radiogenomic features can be explored in the following sequence: RmFs that significantly influence the survival time, either positively or negatively, are selected using our criteria outlined in Section 2.2.3. For each selected RmF, the important RF groups comprising this RmF can be identified by evaluating the sPCA loading information as shown in Figure 2.2. To obtain significant genes and genomic platforms for the selected RmFs, we then trace back to the radiogenomic model and the genomics model, to identify which genes, if any, are associated with that RmF, and then which upstream platforms appear to be modulating the genomic effect. The specific results for each stage are described here.

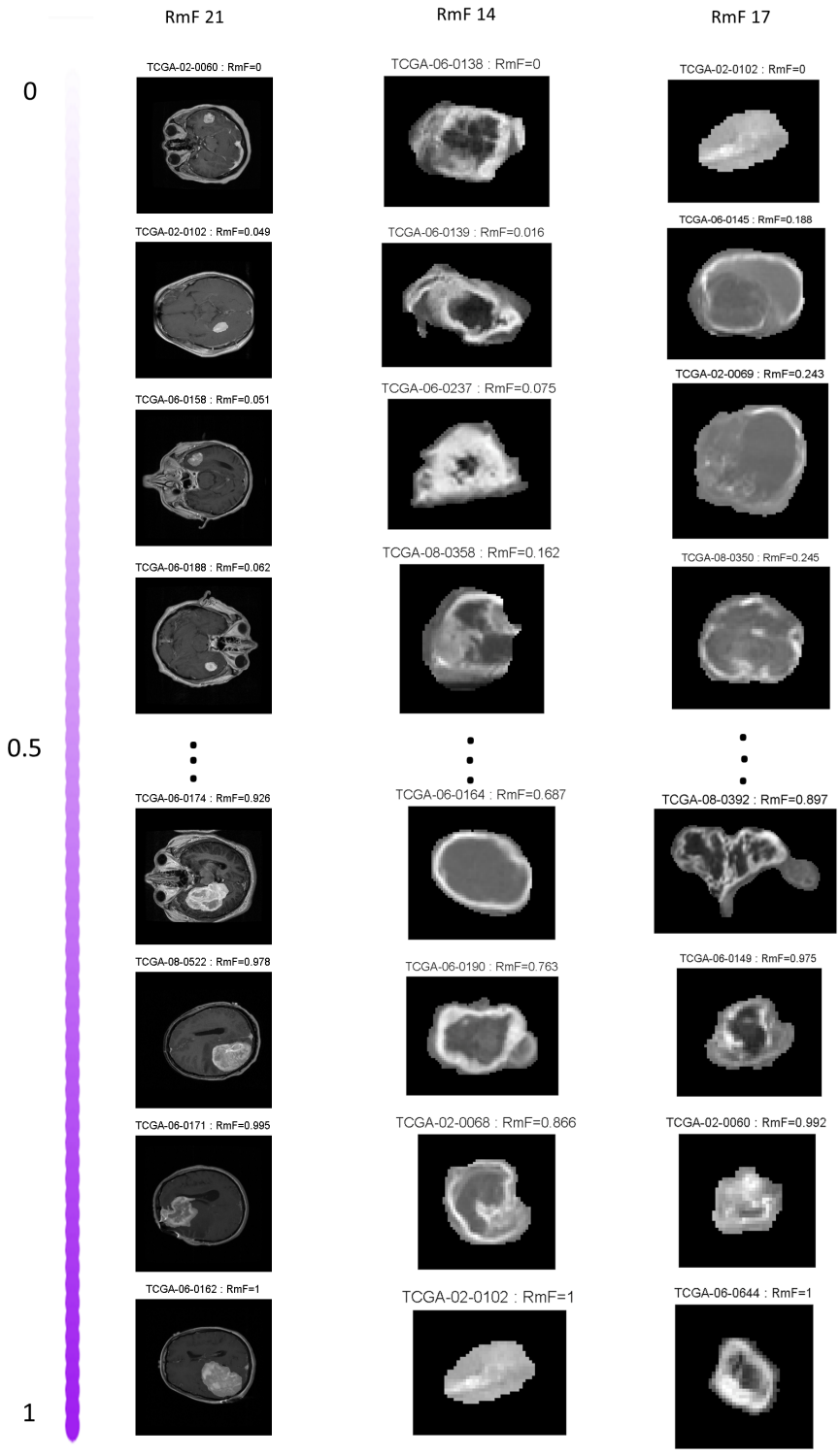


Figure 2.3: T1-Post Contrast images are shown based on the sorted results of 3 representitive RmFs: RmF 21 mainly accounts for tumor area; RmF 14 mainly represents tumor pixel heterogeneity; RmF 17 represents tumor uniformity. The RmF values are all scaled from 0 to 1.

2.3.3.2.1 Radiomic Results

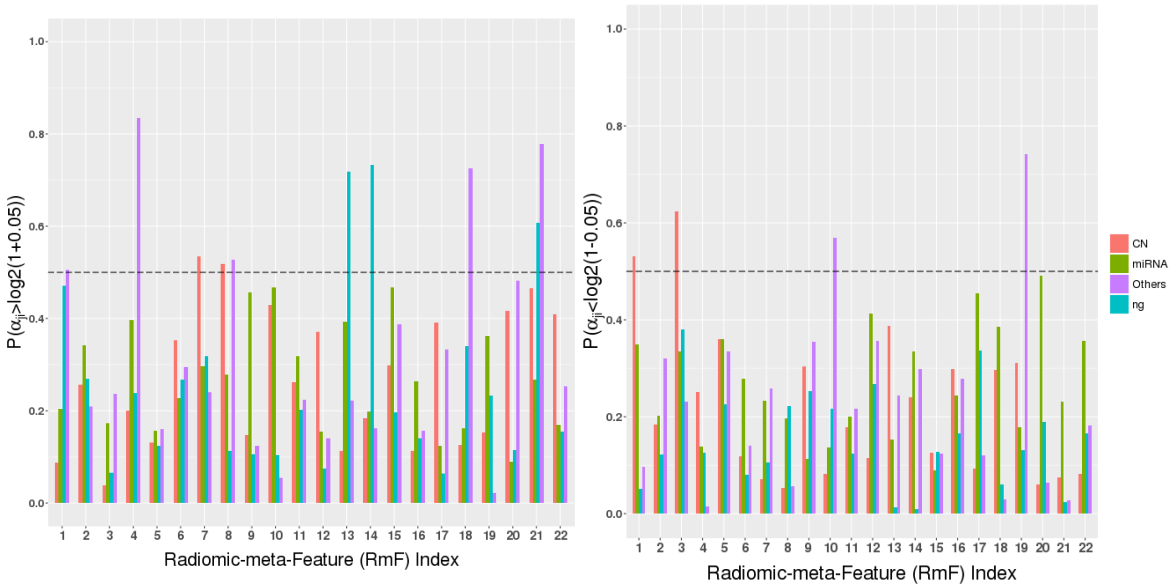
We use posterior probabilities to detect significant radiomic signals as well as genomic platform factors in both stage II and stage III based on the median probability criteria described in Section 2.2.3. Figure 2.4 shows the posterior probabilities used to select the positively and negatively significant clinical RmF combinations. The results show that more RmF are positively significant for the prognostic outcome, with 1 unit change leading to at least 5% increase in survival time ($\delta_2 = 0.05$). Also negatively selected significant RmF combinations have the interpretation as 1 unit change leading to at least 5% decrease in survival time. Based on Figure 2.4, we see that RmF 7 and RmF 8 have a positively significant influence on the survival time, with the parts that are modulated by genes through their copy number effects (G_{CN}). RmF 1 and RmF 3 are negatively associated with survival with the parts that are modulated by genes through their copy number effects (G_{CN}). RmF 1, RmF 4, RmF 8, RmF 18 and RmF 21 are positively related with survival via genemic effects not modulated by CN and miRNA. RmF 10 and RmF 19 are negatively associated with survival through genomic effects not modulated by miRNA nor CN. RmF 13, RmF 14 and RmF 21 are positively associated with survival apart from genetic modulated factors.

To interpret the flagged RmFs, we turn to Figure 2.2, which illustrates how much variance each RF group contributes to the corresponding RmF combinations. RmF 8 is found to be positively associated with survival through CN effects, and Figure 2.2 shows that RmF 8 is dominated by the the RF groups "T1_LoG_Tex_R1" and "T1_LoG_Tex_Fine". RF names and their brief interpretations are shown in the table of Section A.2.2 of Appendix. In general, we see that texture features derived from T1-weighted post contrast images processed with R1 normalizing approach tend to be more significant, and based on the actual loading values, we found Haralick

features to be important, including sum average and inverse difference moment. As another example, RmF 19 modulated by gene expression not explained by miRNA or CN changes ($\boldsymbol{G_O}$) is detected to be negatively associated with survival, and for this RmF the dominant RF group is the Haralick features extracted from T2-weighted FLAIR images, especially with exact features named cluster shade, cluster prominence, energy and contrast. Additionally, RmF 21, which is found to be positively associated with survival both through genomic factors explained by"other" and the non-gene driven part. Further checking found that RmF 21 is associated with T1-weighted post contrast and T2- weighted FLAIR tumor areas. This indicates tumor area, as one of the major regional features, associated with the survival time and seemingly moderated by gene expression of signaling pathway genes, in part regulated by some genomic transcriptional factors other than CN or miRNA.

Radiomic Biological Significance

In general, more radiomic features extracted from T1-weighted post contrast MRI images are selected to be clinically significant and most of them appear to be associated with genomic effects in signaling pathways. This is not unexpected given the fact that recent studies in literature showed that genomics are expected to be most related to T1-weighted post contrast images rather than T2-weighted FLAIR preprocessed ones. More specifically, RmF 14, which mostly captures the contrast margin of the enhancing MRI image, the magnitude and the loading information, indicate that higher texture feature *sum of average* or lower texture feature *sum of variance* derived from the contrast of the edges leads to longer survival times. The detection of RmF 10 shows that histogram features, derived from T2-weighted FLAIR image pixel intensity and representing the global summary of the enhancement, are selected to be primarily affecting patients' survival. It has been shown that the overall in-



(a) positively significant RmF combinations **(b)** negatively significant RmF combinations

Figure 2.4: Results of stage III (*radiogenomic clinical model*): Detecting postively and negatively significant RmF combinations. Each RmF is segmented into 4 parts, of which 3 parts are modulated by different genomic platform combinations denoted as $\boldsymbol{\mathcal{I}_{CN}}, \boldsymbol{\mathcal{I}_{miR}}$, and $\boldsymbol{\mathcal{I}_{O}}$. The 4^{th} part is modulated by unknown/unmeasured factors represented as $\boldsymbol{\mathcal{I}_{\bar{g}}}$ ("ng" in the legend). The barplot shows the posterior probabilities that the coefficient for each part $\alpha_{jk} > \delta_{+}^{*}$, where α_{jk} denotes the k^{th} RmF modulated by the j^{th} genomic platform. For each RmF, the probabilities of these 4 components, CN, miRNA, others, and ng, are respectively shown in red, green, purple and blue. Each probability in Figure 2.4(a) shows that 1 unit increment in the RmF component leads to at least 5% increase in survival time. Each probability in Figure 2.4(b) shows that 1 unit increment in the RmF component leads to at least 5% decrease in survival time. We consider the markers to be significant if this posterior probability is larger than 0.5.

tensity is correlated with blood flow vasoconstriction. Moreover, we see associations with several key genes PDGFRA and TP53, with genomic transcriptional factors that affect the uniformity of the overall pixel intensity. In addition, RmF 21, with both genomic transcriptional factor driven part and non-gene driven part, are also selected to be significant in influencing patients' survival time. Since the region feature, more specifically, tumor area, that captures most of the variation of RmF 21, our conclusion indicates that tumor area calculated from both T1-weighted post contrast and T2-weighted FLAIR images, are clinically important, larger area results in shorter survival times.

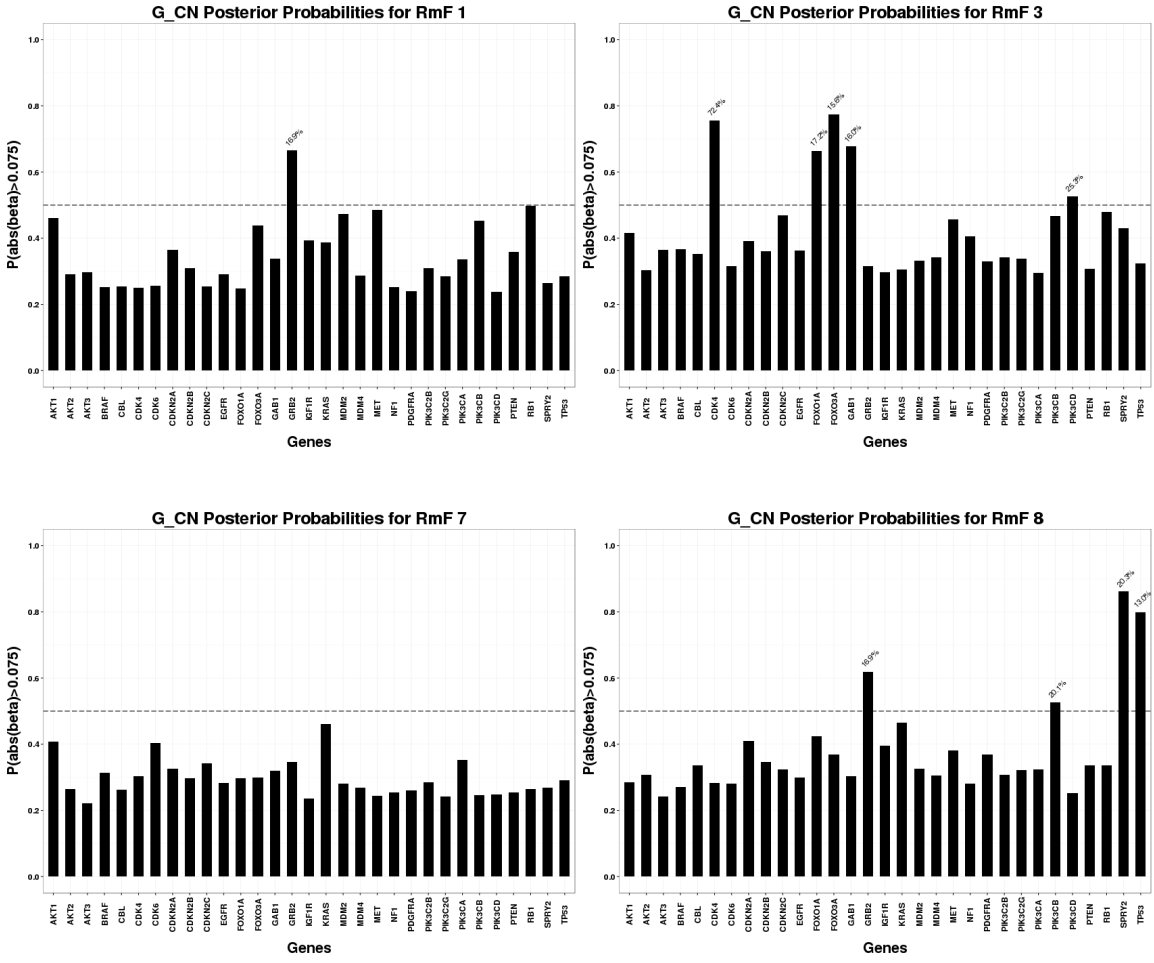


Figure 2.5: Significant genomic CN combinations

2.3.3.2.2 Genomic Results

For the selected RmFs, we trace back to stage II and obtain the regulating genes that significantly affect the RmFs through specific genomic platforms (CN, miRNA or others), as shown in Figure 2.5, Figure 2.6 and Figure 2.7. To flag genes as associated wtih the RmFs, we compute the posterior probabilities of the magnitude exceeding a pre-specified threshold. For our analysis, we present the results with the setting $\delta_1 = 0.075$ in this section since it gave us the best balance between the signal and sparsity. For the flagged genes, we traced back through the stage I model to acquire the percentage values (marked in blue) that represent the proportion of

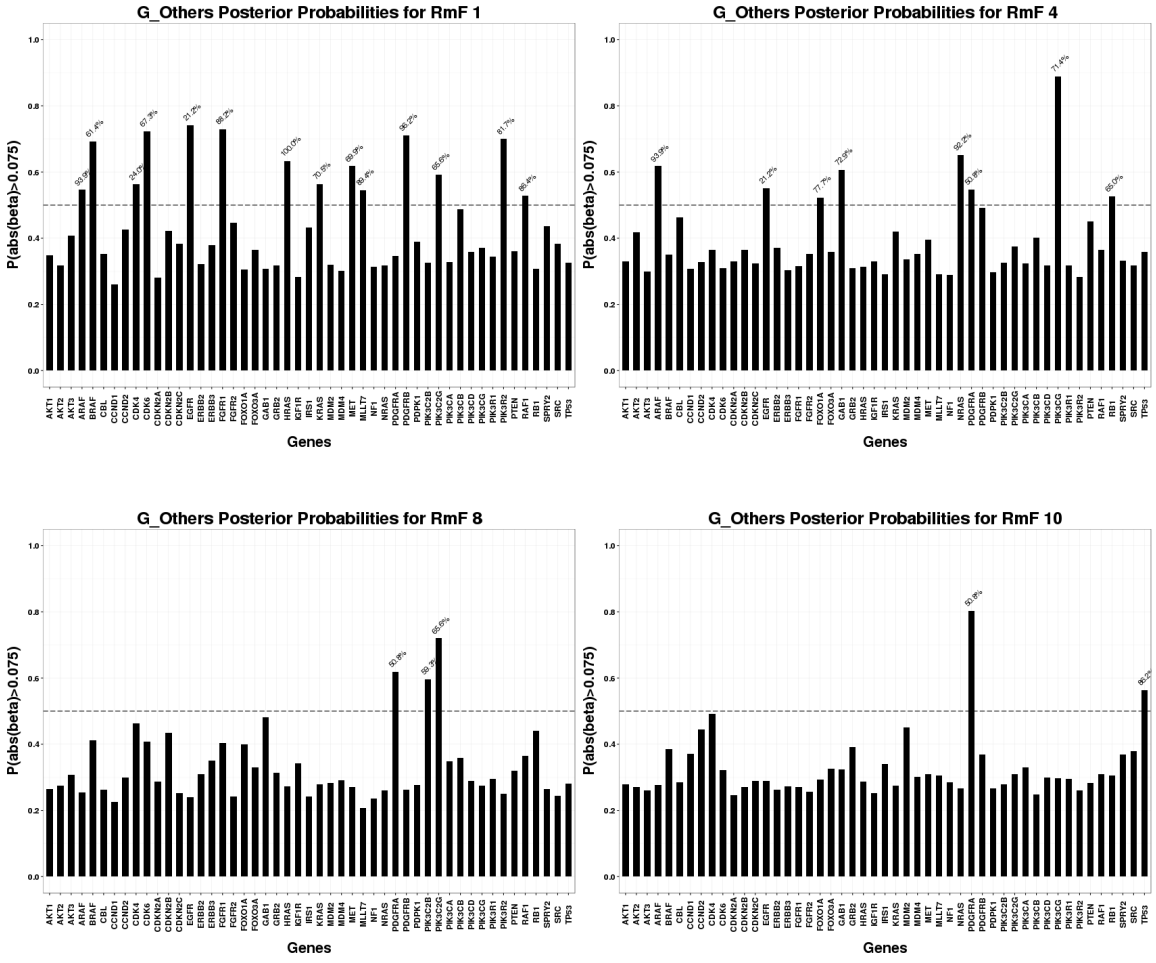


Figure 2.6: Significant genomic mRNA "Other" combinations

the mRNA variance that is explained by the corresponding genomic platform. For example, RmF 8 modulated by the CN combination is selected to be important, referring to the top left graph in Figure 2.5, genes GRB2, PIK3CB, SPRY2 and TP53 are selected as important, affecting RmF 8 through CN alteration. For gene SPRY2, 20.3% of its mRNA is explained by CN alteration. Also, genes PDGFRA and TP53 are selected as significantly influencing RmF 10 via other transcriptional factors. For TP53 in particular, the genomic factors (other than CN and miRNA) explains 86.2% of its mRNA variance.

For the results of stage I, after performing genomic modeling and filtering out

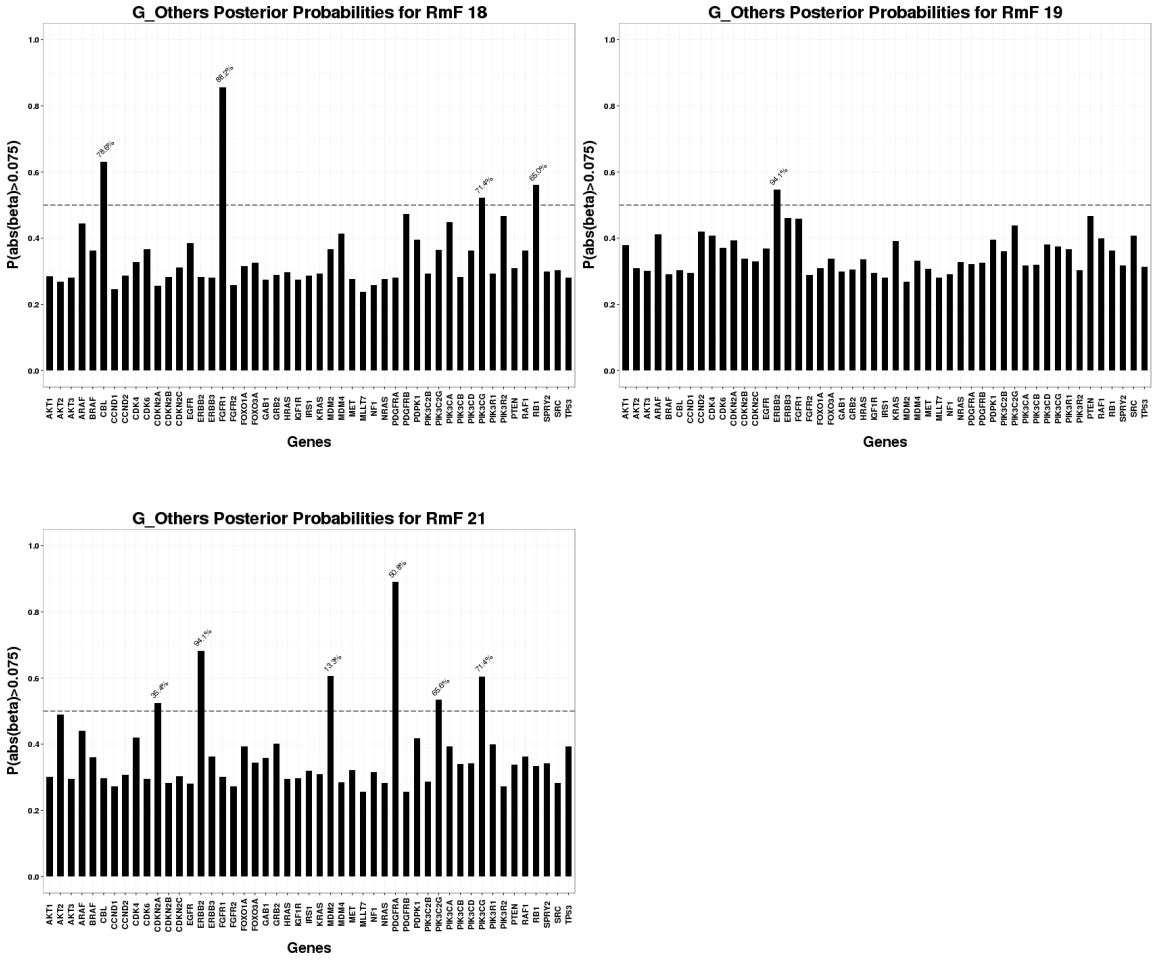


Figure 2.7: Significant genomic mRNA "Other" combinations

the genomic platforms that did not explain much of the variance of gene expression (discussed earlier), there are 92 markers in the remaining gene-platform combinations (miRNA: 12; CN: 31; Others: 49) being carried into stage II, the radiogenomics model, as predictors. Figure 2.8 presents the overall genomic and radiomics results: RmF 7 and RmF 8, modulated by CN, are selected to be positively associated with survival time. Furthermore, 4 genes (GRB2, PIK3CB, SPRY2 and TP53), with their part of gene expression (mRNA) explained by CN alteration, are detected as being significantly associated with these RmFs. For the transcription modulated part, RmF 10 and RmF 19 are detected as being negatively important and associated with gene ERBB2, TP53 and PDGFRA; while RmF 21 is positively significant and associated

with genes CDKN2A, ERBB2, MDM2, PDGFRA, PIK3C2G and PIK3CG. For the non-gene-driven factors, RmF 13, RmF 14 and RmF 21 are positively significant.

RmF Combinations	Selected RmF	Posterior Probability	Magnitude	RF groups	Selected Genes
RmF_CN	**RmF 1**	0.5308	-0.1734	T1_LoG_Tex_R1; T1_LoG_Tex_R2	GRB2(16.9%)
	RmF 3	0.6233	-0.2628	T1_LoG_Tex_R1; T1_LoG_Tex_R2	CDK4(72.4%); FOXO1A(17.2%); FOXO3A(15.6%); GAB1(16.0%); PIK3CD(25.3%)
	RmF 7	0.53485	0.1693	F_LoG_Tex_R1; F_LoG_Tex_R2	-----
	RmF 8	0.5174	0.1393	T1_LoG_Tex_R1; T1_LoG_Tex_Fine	GRB2(16.9%); PIK3CB(20.1%); SPRY2(20.3%); TP53(13.0%)
RmF_Others	RmF 1	0.5053	0.1668	T1_LoG_Tex_R1; T1_LoG_Tex_R2	ARAF(93.9%); BRAF(61.4%); CDK4(24.0%) CDK6(67.3%); EGFR(21.2%); FGFR1(88.2%); HRAS(100%); KRAS(70.5%); MET(69.9%); MLLT7(89.4%); PDGFRB(96.2%); PIK3C2G(65.6%); PIK3R2(81.7%); RAF1(86.4%)
	RmF 4	0.8347	0.4651	T1_LoG_Tex_R1; T1_LoG_Tex_R2; T1_LoG_Tex_Fine	ARAF(93.9%); EGFR(21.2%); FOXO1A(77.7%); GAB1(72.9%); NRAS(92.2%); PDGFRA(50.8%) PIK3CG(71.4%); RB1(65.0%)
	RmF 8	0.5268	0.1340	T1_LoG_Tex_R1; T1_LoG_Tex_Fine	PDGFRA(50.8%); PIK3C2B(59.3%); PIK3C2G(65.6%)
	RmF 10	0.5686	-0.1636	F_LoG_Hist_R1; F_Gauss_Hist_R2	PDGFRA(50.8%); TP53(86.2%)
	RmF 18	0.7263	0.2560	F_LoG_Tex_Fine; F_LoG_Tex_R1	CBL(78.6%); FGFR1(88.2%); PIK3CG(71.4%); RB1(65.0%)
	RmF 19	0.7427	-0.3201	F_LoG_Tex_R1	ERBB2(94.1%)
	RmF 21	0.7776	0.4349	T1_region; F_region	CDKN2A(35.4%); ERBB2(94.1%); MDM2(13.3%); PDGFRA(50.8%); PIK3C2G(65.6%); PIK3CG(71.4%)
RmF_non_gene	RmF 13	0.7187	0.2233	F_LoG_Hist_R1	-----
	RmF 14	0.7327	0.2407	T1_LoG_Tex_R2	
	RmF 21	0.6078	0.1702	T1_region; F_region	

Figure 2.8: Results: Significant RmFs, genes and genomic platforms. Four categories of RmF combinations are listed in the first column, where "non_gene" denotes "$\bar{g}$," which is the non-gene-driven part of the RmF. For each category, several significant RmFs detected from the clinical model are listed in the second column, with unbolded indicating positive ones; bolded indicating negative ones. Posterior probability of the important radiomic markers is shown in column 3. For each selected RmF, several RF groups are selected based on RmF description heatmap (Figure 2.2). For each significant RmF combination, significant genes are listed with the percentage of how much the variance of mRNA is explained by the specific genomic platform.

Genomic Biological Significance

Result table shows that gene EGFR is selected to be significant for multiple flagged RmFs. It agrees with the literature that the aberrations and gene expression of EGFR, with its full name as "epidermal growth factor receptor", have been associated with the classical subtype of GBM among 4 major subtypes (proneural, classical, mesenchymal and neural), defined based on transcription data analysis [110]. This particular subtype accounts for $\sim$ 25%-30% of GBM cases. The amplification of the EGFR gene is the most common genomic change that leads to overexpression of the receptor variant III (EGFRvIII), and 20% or less EGFRvIII in GBM is significantly related to longer overall patient survival [76]. Moreover, PDGFRA is another gene which has been flagged as important for multiple RmFs. It was found that for the proneural subtype, platelet-derived growth factor (PDGF) receptors (PDGFRAs) have been found to represent gene [110]. Also, PDGFR has been positively correlated with patient survival time and its critical role in oncology has been well described in the context of gliomas [80]. Gene TP53 is selected to significantly influence RmF 8 via its mRNA explained by CN, and specifically, TP53 has been found to be the main hub gene that acts as tumor suppressor through comparative analyses of CN and mRNA expression in GBM tumor and xenografts [41]. The study illustrated that loss of TP53 function in GBM leads to transcriptional upregulation in gene expression network.

MDM2 is a commonly known oncogene that inhibits the tumor suppressor TP53; its overexpression and amplification have been studied through the analysis of CN alterations and gene expression profiles in previous studies [118]. The gene CDKN2A, with other transcription factors accounting for its expression, has been found to be significantly associated with tumor area for both T1-weighted post contrast and T2-weighted FLAIR. CDKN2A belongs to the RB1 pathway, serves as

a cyclin-dependent kinase inhibitor, and has been detected to be important [101]. It has been reported that loss of RB1 expression occurs in up to 25% of glioblastomas. Changes in RB1 expression have been associated with alterations in tumor cell proliferation and survival [56], [77]. Also, the assessment of RB1 promoter hypermethylation showed a clear correlation between the loss of RB1 expression and promoter hypermethylation [78]. Analysis of GBM on the molecular level (TCGA data), using fluorescence in situ hybridization and immunohistochemistry, showed that alterations in RB1 occur more commonly in the proneural subtype of GBM.

2.4 Discussion and Conclusion

This Chapter presents the radio-iBAG model, a general framework for multi-scale integrative Bayesian analysis of radiogenomics data. Our hierarchical models incorporate biological mechanistic relationships among multiple genomic platforms, radiomic feature analysis and radiogenomic analysis with relevant clinical outcomes. There are three key features of this modeling strategy: (1) Multiple genomic platform profiles are incorporated in the radiogenomics framework; (2) For model fitting, high dimensionality with a pre-defined group structure in the covariates can be addressed through Bayesian shrinkage priors. In particular, we choose the normal gamma prior due to its flexibility in both shrinkage and parameter estimation; and finally (3) Investigating the relationship between clinical outcomes and radiomic features containing genomic information allows us to identify clinically significant genes, radiomic features and more importantly, the hidden associations between these two data modalities. We note that although our modeling strategy is motivated by an imaging genomics study in GBM, our methodology is general and can be applied to any other disease domain which generates quantitative imaging data with matched genomic data. This includes neurological diseases where the imag-

ing features could be computed from structural or functional neuroimaging assays [7].

We applied our methodology to the analysis of radiomic and genomic data sets of GBM. Our model analyzed the relationship between the survival times of patients and the RmFs modulated by various gene-platform combinations. Our analysis identified several RmFs that significantly impact survival times as well as identified the key radiomic features driving these factors. These results revealed that some of the most prognostically important radiomic features include tumor area, intensity histogram uniformity, and Haralick features derived from the GLCM, including energy contrast, inverse difference moment, and entropy for both T1-weighted post-contrast and T2-weighted flair images. Based on the results of modeling the relationship between RmFs and multi-platform genomic measurements, for each detected RmF, we subsequently identified which gene-platform combinations modulated that RmF. This allows us to detect prognostic RmFs modulated by upstream molecular platforms such as copy number, microRNA or other factors. Furthermore, we were able to identify which genes and platforms were driving these associations.

In summary, the advantages of applying integrative analysis of multiplatform genomic profiles in this framework are illustrated through the hierarchical backtracking, which allows us to discover strong associations and interrelationships among the clinical, image, and genomic factors that may help elucidate the underlying biology. Most of the significant genes identified in our analysis have been shown to be biologically and clinically relevant to GBM molecular subclassifications, cancer development, or therapeutic strategies.

The Radio-iBAG model that we proposed has several limitations. First, from clinical perspective, the model is more of an exploratory analyses that mainly inves-

tigate biological regulatory mechanisms, thus, further work would have to be done to establish its translational relevance. Also, the multi-level model that we construct in this chapter is complex and takes careful thought to interpret. Another limitation lies in that the model does not include clinical features especially the ones such as age and Karnofsky score, that have been shown to have significant impact on survival times of patients with GBM [86]. In addition, from the aspect of result validation, we currently apply our method into GBM as case study merely based on the data sets that come from one resource (TCGA), the results could be further validated from other data resources, which will quantitatively lead to more consolidated detection of genomic and radiomic markers.

Several possible future extensions and generalizations could be explored based on our Radio-iBAG framework. For example, in our methodology, we applied a multi-stage modeling strategy in doing integrative analysis. A possible advancement may be using a joint model to capture all the relationships among different platforms simultaneously and maintain the detective power with interpretable results. One other possible direction may be incorporating pathway information as another hierarchy into the model structure or considering more complicated biological mechanisms at molecular level, e.g, hidden associations between a gene and other platforms of the neighboring genes, into the modeling framework, e.g. as considered in [71]. Another possible future extension may be involving histological images of different tumor tissue regions as another imaging modality into the study, which will provide more pathological based interpretable radiogenomic relationships along with relevant clinical outcomes. We leave these tasks for future work.

Chapter 3

pathDrive: identification of pathway-specific upstream genetic and epigenetic drivers

3.1 Introduction

Cancer is characterized by inter- and intra-tumoral heterogeneity, it occurs when a group of cells experience a series of molecular alterations enabling them to behave outside of the control of normal bodily functions. These changes give groups of tumor cells the ability to grow uncontrollably, metabolize energy in various ways, evade immune attack, and in advanced cancers the ability to invade neighboring tissue regions, gain mobility to travel around the body, build and maintain their own blood supply, and acquire stem cell-like abilities that make them highly adaptable and equip them with survival and growth advantages over other non-cancerous cells. While all cancers share some combination of these key hallmarks, each patient's cancer is characterized by a unique combination of underlying molecular alterations, making cancer inherently heterogeneous and thus difficult to diagnose and treat,

as treatments that work for one patient may not work for others. The challenges drive current modern cancer research in developing precision medicine which targets to individualized treatment that involves the application of new diagnostics and therapeutics, aimed to the necessities of a patient based on his or her own molecular characteristics [47].

In recent years, many advanced innovative technologies are rapidly enhancing and expanding the field of precision medicine. These include microarrays and next-generation sequencing methods that can measure mRNA expression levels, DNA mutations and copy number changes on a genome-wide scale, as well as proteomic techniques to measure protein abundances. Given these advanced technologies, more systematic picture of genomic and epigenomic regulatory mechanisms has been provided by large-scale and multi-omics profiling of cancers. Furthermore, genomic and epigenomic data sets such as copy number alteration, mutation, mRNA, methylation, histone modification and microRNA data sets are all publically accessible via The Cancer Genome Atlas (TCGA) [108] with advanced tools that have been developed in quantifying epigenetic and transcriptional factors on a genome-wide scale. It all leads to the current rapid-developing area of integrative analysis which targets to analyze heterogeneous types of data from inter-platform technologies, delineating complicated biological mechanisms at molecular level.

The advent of high-throughput sequencing and profiling technologies has led to the formation of lists of differentially expressed genes and proteins, leaving new challenge in extracting main biological meaning from the long lists. Driven by this particular challenge, researchers focus on partitioning the long lists of individual genes into smaller groups of related genes or proteins that function in the same pathways. The pathway gene sets analysis alleviates the complexity in analyzing

thousands of individual genes or proteins, and more over, it leads to better underlying biological interpretation in identifying active pathways with particular functional processes, biological components or structures [54]. Therefore, modern analysis tends to focus on pathways which can be regarded as more practical units to obtain functional biological study, given that the silencing of a key gene may not have a functional impact given the ability of the body to bypass certain molecules and activate compensatory mechanisms that can achieve the same functional result using an alternative molecular mechanism. Furthermore, given pathway analyzing and the advanced sequencing tools, regulation of genes at pathway level that carries out cellular functions has been recently investigated. Wilk et al., 2017 [116] identifies specific associations at the systems level in cancer through integrative analysis of miRNA and overall behavior of the pathways that show functional consequences of miRNA dysregulation. The trend leads to the study of genetic and epigenetic regulatory factors that explain inter-patient heterogeneity in particular pathway activity, elucidating functional biological mechanisms in the molecular processes underlying specific pathways.

In this Chapter, we develop the pathDrive modeling framework in which we compute patient-specific measurements of activity for a given pathway from gene expression and then identify a small number of upstream genetic and epigenetic regulatory factors that explain the pathway activity. Our main goal is to identify sparse set of upstream factors that can be regarded as key leverage points or switches that drive functional activity of a pathway. The selected regulators can be further explored as potential molecular targets for precision therapy. In this Chapter, we will describe the general pathDrive model framework outlining how we define and measure pathway activity levels, methods that we obtain the matched upstream factors of multiple genomic platforms and the modeling algorithms. In our case study,

we apply our framework in identifying potential therapeutic targets in Colorectal Cancer incorporating its Consensus Moleculary Subtypes (CMS) [33] so as to get CMS-specific therapeutic targets.

In Section 3.2, we introduce our general methodology flow in pathway score calculation, upstream factor mapping as well as the modeling framework. In Section 3.3, we illustrate our proposed method on Colorectal Cancer (CRC) CMS-specific pathway analysis with detailed description of the key pathways characterizing specific CMS biology and the detection of the genomic and epigenetic switches. Section 3.4 shows the main results and biological interpretation. In Section 3.5, we draw the conclusions and discussion on some future extensions and advancements.

3.2 Method

In this thesis Chapter, we develop the pathDrive modeling framework in which we compute patient-specific measurements of activity for a given pathway from gene expression and then identify a small number of upstream genetic and epigenetic regulatory factors that explain inter-patient heterogeneity in this pathway activity. In this way, we utilize genetic pathway analysis to characterize large-scale functional activity of a pre-specified molecular process, and then find regulatory factors upstream to gene expression that are key "leverage points", or genetic and epigenetic switches that may be drivers of the functional activity. We will describe the general pathDrive model, outline how we calculate the pathway activity score, how to collect the upstream factors as covariates and the modeling strategy that is used in our framework to identify potential upstream factors.

3.2.1 Pathway Score

Gene set enrichment (GSE) analysis provides a framework that focuses on pathway signature rather than individual gene expression via the generation of enrichment scores. Current methods in this framework can be generally partitioned into supervised and unsupervised, population and single sample measurements. GSEA approach described in Tian et al., 2005 [106] is supervised and population based method that calculates the mean expression for a specific set of genes compared with those that are out of set. Other relative supervised methods include EPSA [105], expression pattern analysis [79] and etc. However, in our analyzing framework, the pathway score needs to be unsupervised and single sample based with specific summary of the gene pathway activity level for each patient. There are various algorithms that belong to this category. Pathway Level analysis of Gene Expression (PLAGE) [109] that utilizes Singular Vector Decomposition (SVD) to obtain a 'pathway activity level' that is calculated based on the expression levels of the genes in the pathway. Also, ssGSEA is another sample-wise enrichment score calculated through the difference between Empirical Cumulative Distribution Functions (ECDF) of rank-normalized gene expression of the genes inside and outside the specific pathway [8]. Combined z-score method computes the mean of the z-score across all genes within a pathway for each sample after normalization [62]. A recent study introduced another sample-wise nonparametric unsupervised algorithm called Gene Set Variation Analysis (GSVA) [35]. This type of pathway score is measured as a function of the rank-based gene expression of the genes inside and outside the pathway. According to the study and the simulation results in Hanzelmann et al., 2013 [35], GSVA, compared with other relative methods, yields in several advantages such as higher power in detecting subtle differential pathway activities across all samples and higher survival predictive accuracy.

Thus, in this Chapter, we chose to apply GSVA as our pathway activity measurement as outcomes, which can be directly calculated from mRNA matrix with publically available R package named "GSVA" at `http://www.bioconductor.org/packages/release/bioc/html/GSVA.html`. However, the outcome of our pathDrive analysis can be applied using any subject-specific pathway score based on biological context and purpose.

3.2.2 Genomic and Epigenomic Factor Mapping

A crucial aspect of pathDrive is specification of an appropriate set of potential upstream genetic and epigenetic regulatory factors. For a given pathway, we consider potential upstream regulatory factors summaries of mutation, copy number, and CpG methylation for any genes present in the pathway, plus expression levels of any miRNA or transcription factors (TF) associated with any pathway genes. Therefore, our multiplatform genomic data sets are expressed as data matrices: X_{mRNA}, X_{CN}, X_{miRNA} or X_{Methy} and $X_{Mutation}$, each with rows as samples and columns as genomic features (either gene level summaries or matched sites) of the respective platforms.

The terms in the model are described and interpreted in details as follows, for gene g,

- X_{miRNA_g} is an aggregated miRNA expression value matrix that integrates information across microRNAs that have been documented to regulate the expression of gene g. For a given gene, there exists multiple microRNAs that interact with this gene. Several mRNA-miRNA membership matrices provides us the mapping references, e.g. targetHub [69], which collected miRNA-mRNA interaction records based on 5 external databases, multiMiR [69] is generated based on 14 external databases, including validation databases, prediction databases and

drug-associated databases. Here in our thesis Chapter, we apply targetHub in our mRNA-miRNA mapping, and for a specific gene, we include all the mapped microRNAs into covariate matrix that we incorporate into our model.

- X_{CN_g} are gene-level summaries of the CN alteration for the g^{th} gene with dimension $N_{\mathcal{G}} \times CN_g$. Similarly, as there are multiple CN alteration values from different markers within the same gene, CN_g denotes the number of gene-level vectors. In our model, we include copy number alterations that are measured as the average copy number within the specific gene region.

- X_{Methy_g} contains matched CpG sites values for gene g. Since simple gene-level summaries such as mean methylation across all CpG results in limited biological interpretation, here we applied Gene-Specific Methylation Profiles (GSMP) devloped in Liu et al., 2017 (submitted) which identifies a small tissue-specific set of functionally relevant CpGs (typically 5-10) whose methylation values significantly predict expression for each gene in the genome. It considers the complexity of the measurement of methylation given the following biological factors: (1) CpGs within gene body or up to 500kb upstream or downstream of the gene can be relevant described in Aran et al., 2013 [5]; (2) there are a large number of potentially relevant CpG sites per gene (median across genomes of 248/gene on 450k methylation arrays, referred to Liu et al., 2017 (submitted)); (3) only a small subset of CpGs are expected to be functionally relevant in particular gene/tissue context; (4) which CpG are functionally important varies across tissue types [53]. Thus, our pathDrive model includes the potential methylation beta values for any CpG in the GSMPs of any pathway genes or associated transcription factors as part of the upstream factor covariates.

- $X_{Mutation_g}$ is the mutation of gene g, which contains summaries of mutation (1=mutation in gene body, 0 none). To obtain sufficient information, we filter out the mutations which have smaller than 2% frequency.

3.2.3 pathDrive Modeling

For each pathway, suppose the GSVA score values for n subjects can be expressed as Y with the dimension $n \times 1$ and the pathway gene set contains m genes, each of the element $Y_i \in [-1, 1]$. For each gene, we match corresponding different platform data sets as described above. The upstream factor matrices were then combined in categories, denoted as : $X = \{X_{CN}, X_{Methy}, X_{miRNA}, X_{Mutation}\}$, which forms the entire genomic upstream factor matrix that is regarded as covariate/design matrix in the modeling part.

Thus, to quantify the relationship between pathway score and upstream regulatory factors, we consider linear regression model that can be expressed as $Y = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}\sigma^2)$. Given the high dimensionality and the assumption of the sparsity, a proper feature selection algorithm need to be selected, which leads to high prediction accuracy and the interpretability of regression model. There are different variable selection methods such as Ridge Regression [43] which contains l_2-norm regularization term in the linear regression estimation, LASSO (least absolute shrinkage and selection operator) [107] which contains l_1-norm regularization term, Elastic Net [124] which can be viewed as the combination of ridge regression and LASSO. Other Bayesian shrinkage methods in variable selection can also be considered, such as spike-slab prior which has been widely applied with good theoretical properties [52] [98], Bayesian shrinkage methods with Global-local priors including normal gamma prior [88], Dirichlet Laplace [13] and etc. Technically, we can apply any of the methods for variable selection, however, considering about the simplicity

and efficiency, we chose to apply LASSO that performs both variable selection and regularization. Moreover, considering the consistency of variable selection, at the same time, we incorporate stability selection as developed by Meinshausen, Nicolai and Bühlmann, Peter [73]. The stability selection approach imbeds sub-sampling technique into high dimensional variable selection algorithms as in our context LASSO, which has been proved to have several attractive properties in Meinshausen et al., 2010 [73], such as selection consistency, depends little on the initial regularization parameters, and better multiple -testing error control.

More specifically, at each stability selection iteration, a random subsample of $1, 2, 3, ..., n$ of size $\frac{n}{2}$ is drawn from the entire sample set without replacement. Then, LASSO modeling is processed based on the subsample set, that is, regularization parameter λ is tuned via cross validation and the final λ together with the selected features are extracted. We repeat the same process for S times and calculate the selection probability for each covariate. The ones with probability larger than 50% are regarded as significant in affecting the pathway activity. Our final linear regression model will be estimated only using the significant features with the whole sample set, and the R^2 is computed for model evaluation. Moreover, as we also target to evaluate the performance using stability Lasso, we measure the concordance correlation coefficient [60] based on 5-fold cross validation across the entire samples. We apply the methodology and evaluation framework for each pathway gene sets that we target, and summarize the results including evaluation measurements such as R^2 and concordance correlation coefficients, and the selected pathway drivers.

3.3 Application

Our methodology can be broadly applicable to any cancer types, but it is initially applied to two important gastrointestinal cancer types: Colon (COAD) and rectal (READ) cancers. As we combine these two cancer types into one group, it has been defined as colorectal cancer (CRC). Moreover, the biological information that can be incorporated into the analysis is Colorectal Cancer Consensus Molecular Subtypes (CMS) that is studied in Guinney et al. 2015 [33]. Our goal is to identify genomic and epigenetic upstream factors that can be regarded as potential therapeutic targets for key pathways that characterize CMS subgroups. We discuss pathway selection, CMS groups and modeling results in our application section.

3.3.1 CMS groups

Colorectal cancer is the third most common cancer type and a leading cause of cancer death worldwide. Guinney, et al., 2015 [33] assembled the colorectal subtyping consortium (CRCSC) and identifies consensus molecular subtypes (CMS). This was done by assembling a data base of gene expression measurements from 4,151 CRC patients from an international collection of 18 studies, having each of the six subtyping systems applied to each of these samples, and then using a network analysis to identify consensus clusters. From this analysis, we introduced a system of four consensus subtypes with distinct biological characteristics: CMS1 (Immune, 17.5%) is characterized by enrichment in hypermutation, hypermethylation, microsatellite instability (MSI), and immune activation. CMS2 (Canonical, 42%) demonstrates canonical CRC characteristics including epithelial differentiation, MYC and WNT activation, and high levels of chromosomal instability. CMS3 (Metabolic, 13%) is epithelial and showed high levels of metabolic dysregulation, with higher rates of KRAS mutation. CMS4 (Mesenchymal, 27.5%) shows characteristics of epithelial

mesenchymal transition (EMT), activation of TGF-, angiogenesis, and prominent reactive stroma. Thus, we incorporate this information into the analysis of pathDrive in CRC, detecting important pathways with their activity significantly differentiating the 4 subtypes and delineating the pathway molecular switches that influence the selected pathway activity.

3.3.2 CRC validation data sets

Three gene expression data sets are used to obtain the commonly important pathways, which are: CRCSC Gene Expression Data, MDACC Integromics Colorectal Cancer Data and TCGA Colorectal Cancer Data. Using these three data sets, we pre-select the important pathways with their scores that significantly differentiate the CMS groups as described above. Below we further describe the data sets in details.

CRCSC Gene Expression Data

This is a gene expression data set from Affymetrix arrays from 1500 subjects and 13 studies that were part of the CRCSC data set.

MDACC Integromics Colorectal Cancer Data

The in house clinical collaborators have routinely banked primary tissue samples from M.D. Anderson patients and obtained a similar panel of multi-platform genomic measurements as those that are collected by TCGA. The current data set consists of 261 primary colorectal cancer resection speciments, collected from 2001-2009 and predominantly stage 2 and 3 with some stage 4. All samples had frozen tissue stored and gene expression measured by Agilent microarrays used as part of the CRCSC cohort. Additionally, these samples have been subjected to

intensive interrogation through a variety of genomic platforms including whole exome DNA sequencing, RNAseq, copy number measurements, 450k methylation arrays, miRNA and other non-coding RNA, and RPPA protein arrays, amongst others. These have been processed in house by faculty in the Department of Bioinformatics and Computational Biology in an analogous manner as the TCGA data.

TCGA Colorectal Cancer Data

The Cancer Genome Atlas (TCGA) is a public cancer genomic data set, derived from a project, aiming to provide more comprehensive information of human cancer genomes by creating an "atlas" of high-throughput multiple genomic profiles across multiple cancers, was launched in 2005 as a publicly funded project [108]. This worldwide initiative involves the study of 32 cancers, and each cancer has on the order of several hundreds of samples for which each genomic, epigenetic, and proteomic platform was implemented. Samples were preprocessed and batch corrected in a common way in Genome Data Analysis Centers, of which M.D. Anderson was one, and the data was made publicly available through in house portal "http://bioinformatics.mdanderson.org/TCGA/gsresults/". Colon (COAD) and rectal (READ) cancers were two of the types, and between the two of them we have a total of 604 patient samples with whole exome mutation, copy number, 450k methylation array, RNAseq, miRNA panel, and reverse phase protein arrays (RPPA) measurements.

3.3.3 Target Pathway Gene Sets

In our analysis, the pathway gene sets were browsed and downloaded from Molecular signatures database (MSigDB) [67], which is a public online resource that contains the largest number of annotated gene set collections for Gene Set Enrichment

Analysis (GSEA) [67] with the website `http://www.broadinstitute.org/msigdb`. MSigDB contains 8 major collections based on established biological processes, states or cancer oriented gene expression studies.

We chose to focus on the following gene sets which are typically studied in the GSEA literature. The pathway gene sets collections that we focused on are *Biocarta, Genomic_Locus, Hallmark_Collection, KEGG_Pathways, Oncogenic_Signatures, Reactome_Pathways*. The number of the member gene sets within each collections are listed as follows: BIOCARTA: 217, HALLMARK: 50, KEGG: 302; Oncogenic_Signatures: 189; REACTOME: 674; Genomic_Locus: 326. The membership genes matching to each gene set could also be downloaded from MSigDB [67], which could be further analyzed both for pathway analysis or search for matched upstream factors.

With the computed GSVA pathway scores for each pathway and each subject respectively from MDACC, TCGA, and CRCSC data sets, our "target pathways" were extracted based on the criterion that if the pathway could significantly differentiate the consensus molecular subtypes (CMS) of colorectal cancer according to the calculated statistical tests. Thus, the pathway is defined being characteristic if it satisfies the following criteria:

- The ANOVA p-value of the GSVA score for comparison across all CMS groups is significant after adjusting for FDR ($FDR < 0.05$)
- For that CMS, the mean GSVA score is either highest (characteristic high) or lowest (characteristic low) across the 4 CMS.
- The pairwise GSVA comparison of that CMS group vs. all 3 other groups is small (< 0.05 unadjusted)

The final list of the target pathway gene sets are the ones which passed the above criterion with at least 2 data sets (MDACC, TCGA and CRCSC Affy). Our pathDrive model will only be established with these pathways in our further analysis.

3.4 Results

3.4.1 Overall Result

Based on the criteria we select important pathways, below we summarize the number of the targeted pathways with the pathway score significantly remain the largest and smallest respectively for CMS1, CMS2, CMS3 and CMS4. More specifically, for example, for CMS1, we calculate the number of the pathways that have pathway score the highest and the lowest with significant difference for CMS1 group compared with others and significantly different in pairwise comparison as well. The following Figure 3.1 shows the corresponding summary. It illustrates that CMS4 group has the largest number of the important pathways, while CMS3 has the smallest.

CMS Group	Min	Max
1	49	253
2	145	46
3	71	54
4	292	381

Figure 3.1: Summary of the number of the targeted pathways for each CMS group

We apply the methodology framework to each important pathway for CMS groups, and overall, we achieve comparatively sparse set of the predictors that are selected to significantly affect the pathway score. Figure 3.2 illustrates the percentage of the predictors that are selected for each pathway gene sets respectively for each CMS

group with histogram. The right skewness and the small value of the percentage indicate sparsity with CMS3 pathways selecting more sparse upstream factor subsets. And the percentage for the entire pathways is also shown in Figure 3.3. Given the sparsity, we show that the variance of the pathway activities that are explained by the spase sets are in general high enough to implicit the parsimonious, that is, we calculate the R squared for training set (entire data set) and Figure 3.4 shows its histogram. Moreover, we do 5-fold cross validation to test the stability and compute the concordance correlation coefficients for each pathway and the histogram is shown in Figure 3.5. Thus, it shows that for some pathways, our methodology successfully achieved our goal in detecting the "leverage points" that explain enough variance of the pathway activities. And we further explain the results for some specific targeted pathways in details with biological interpretation in Section 3.4.2.

Among the selected genetic and epigenetic factors, the proportion of different platform respectively for each CMS group is shown by pie-chart in Figure 3.6. The results show that Methylation and microRNA in general, comprise the large proportion of the selected factors, which is reasonable given our construction of the potential upstream factors, that is one gene has one corresponding copy number alteration and one mutation status, however, it can be mapped with multiple microRNA and Methylation CpG sites. The percentages for overall pathway gene sets are shown in Figure 3.7.

3.4.2 Specific Results and Biological Interpretation

Here in this section, we extract some modeling results of some typical and representative pathway gene sets respectively for CMS groups. We summarize the results from the following aspects: number of the genes in the pathway, heatmap of the gene expression with pathway scores, modeling results with visualized plots and biological

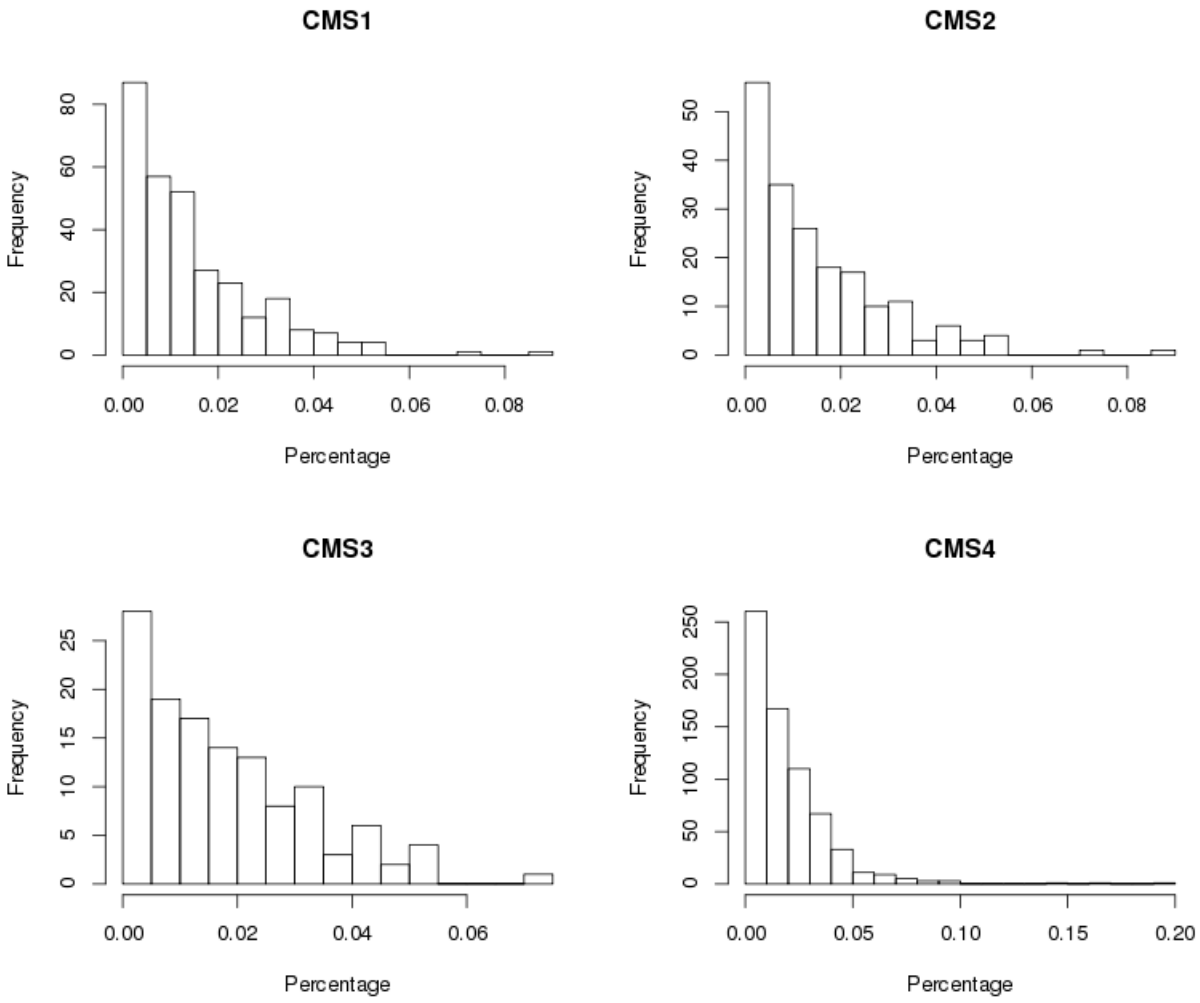


Figure 3.2: Histogram of the percentage of the number of the selected predictors for each CMS group based pathways

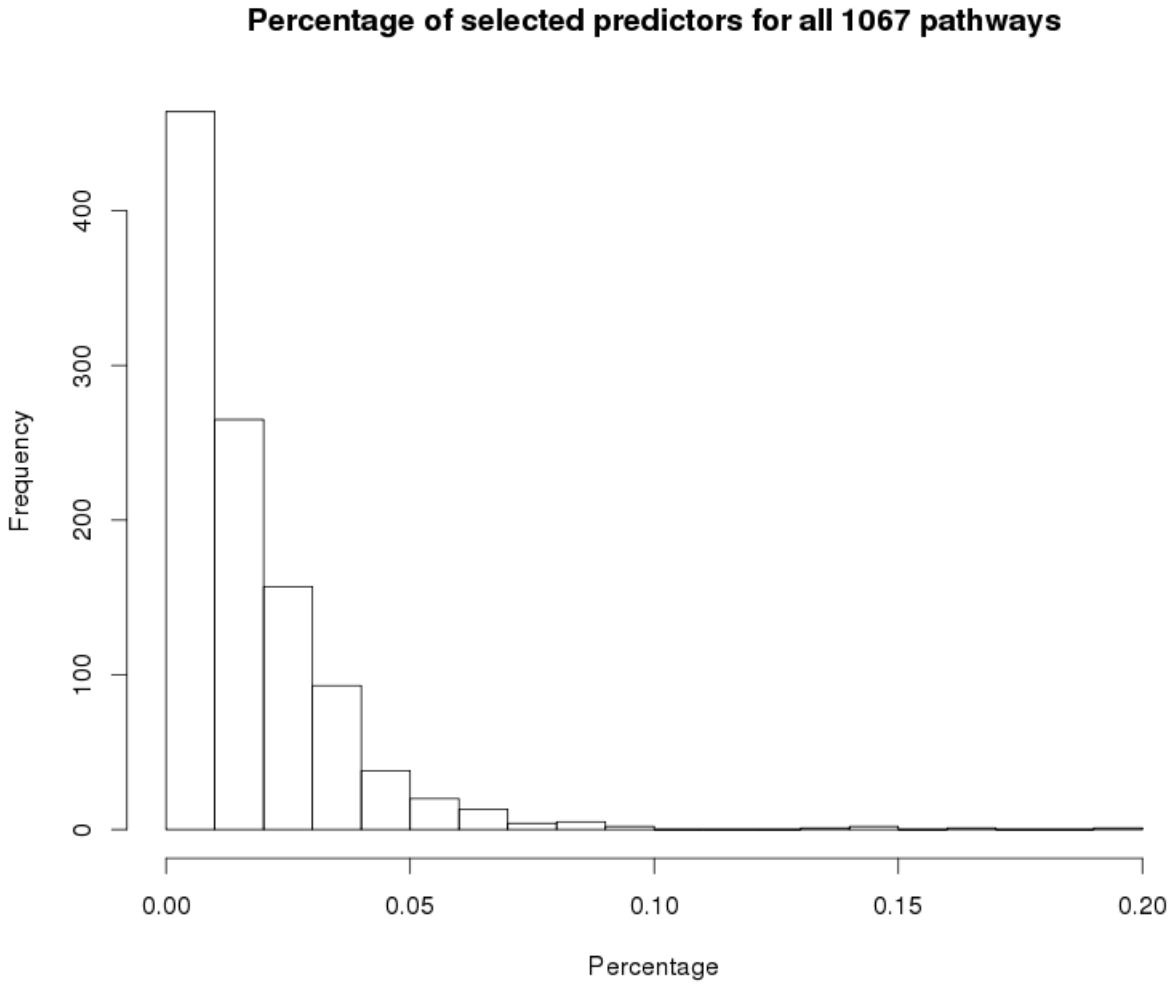


Figure 3.3: Histogram of the percentage of the number of the selected predictors for all targeted pathways

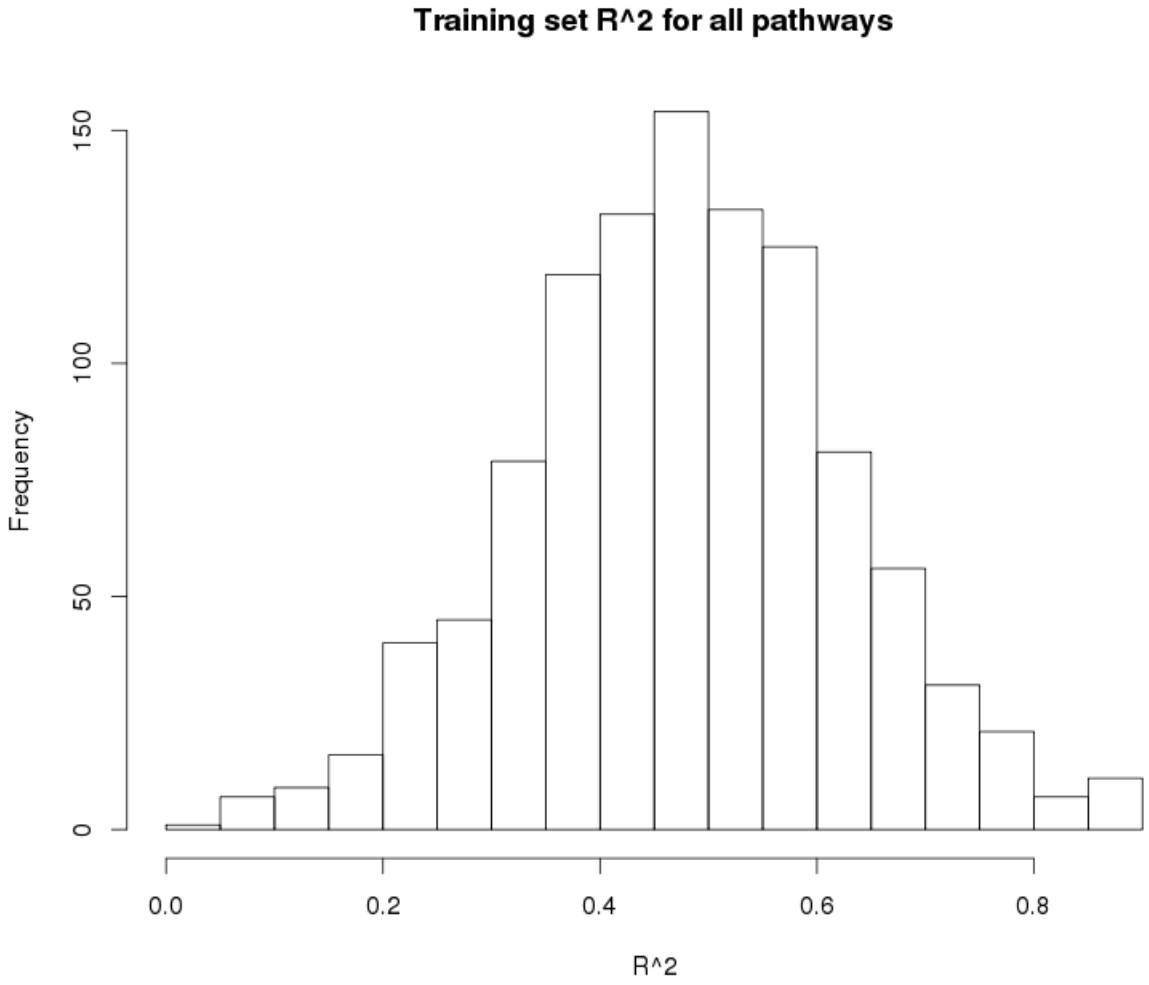


Figure 3.4: Histogram of the R^2 estimated by the selected features of pathDrive modeling for all pathways

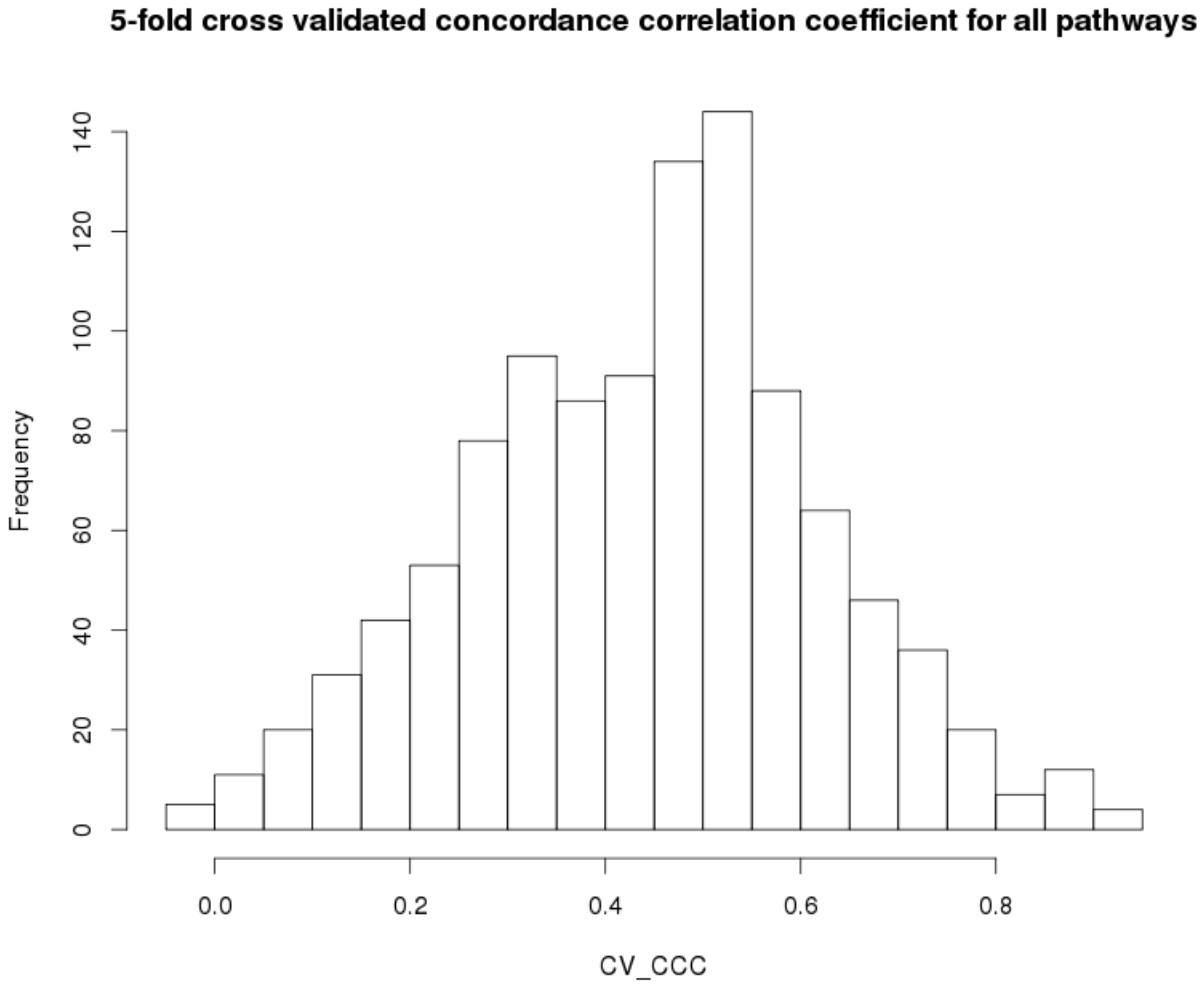


Figure 3.5: Histogram of the concordance correlation coefficients estimated by the selected features of pathDrive modeling for all pathways

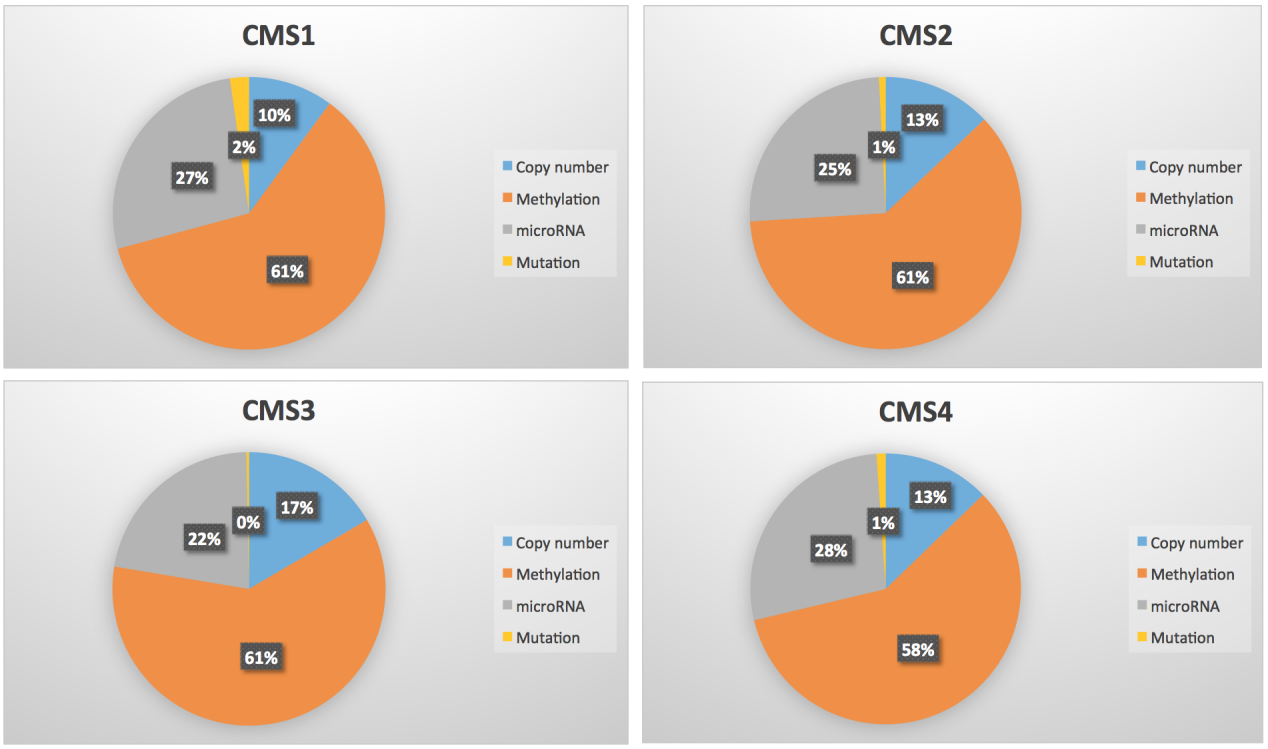


Figure 3.6: Pi-chart of the percentage of the upstream factors for each platform and for each CMS related pathways

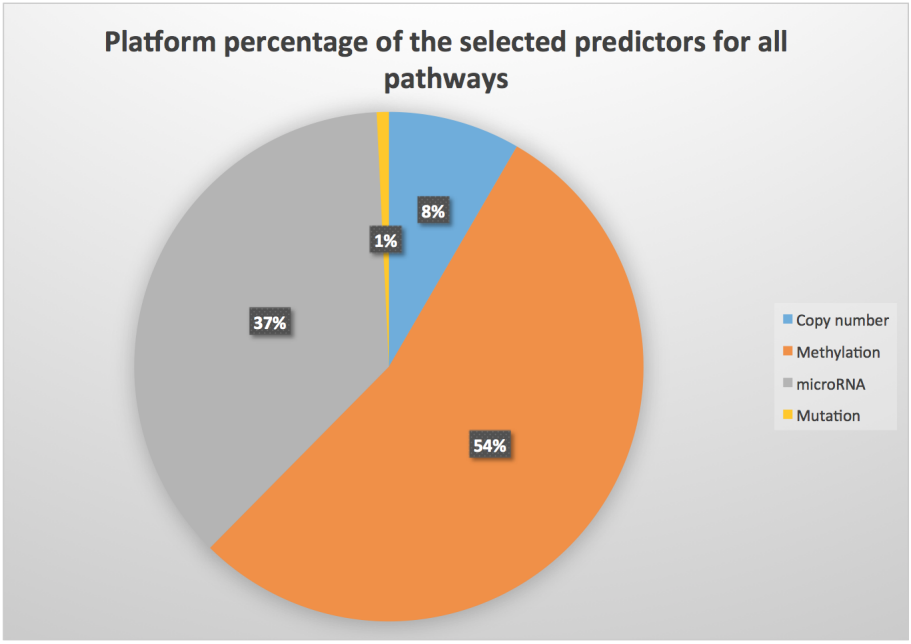


Figure 3.7: Pi-chart of the percentage of the upstream factors for each platform and for all targeted pathways

interpretation.

REACTOME_INTERFERON_GAMMA_SIGNALING

Results:

This signaling pathway belongs to REACTOME collection with 59 membership genes, it is selected as the important pathway that significant has the highest pathway activity in CMS1 compared with that of other groups, which has been validated by our 3 data sets (CRCSC Affy, TCGA and MDACC), as shown by the boxplot in below.

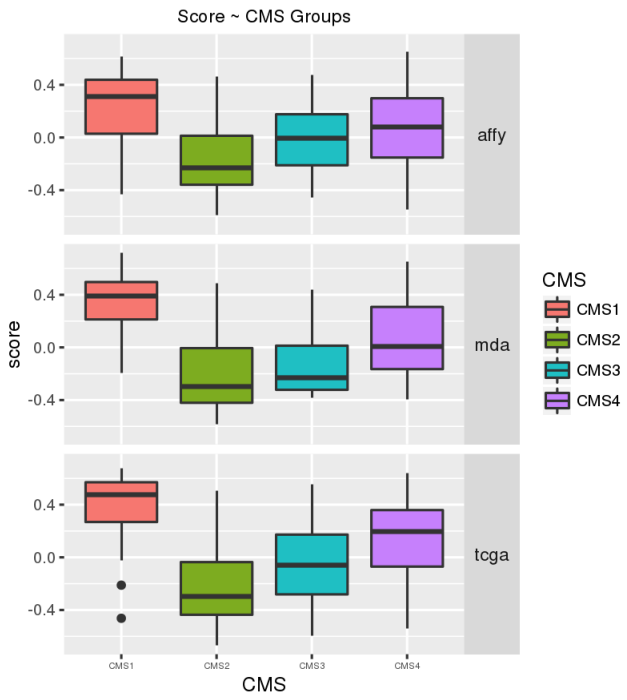


Figure 3.8: Boxplot of GSVA scores across 3 datasets and across CMS groups (REACTOME_INTERFERON_GAMMA_SIGNALING)

The mRNA heatmap with samples sorted by GSVA score and clustered by genes illustrated in Figure 3.9 provides us visualized perspective of the score measurement of the pathway activity, it is clear that the ones with higher GSVA score, in general, tend to have higher mRNA values for most of the genes. Moreover, based on the CMS group label, we can conclude that CMS2 subpopulation tend to have low pathway activity while the CMS1 subpopulation having comparatively higher pathway activity.

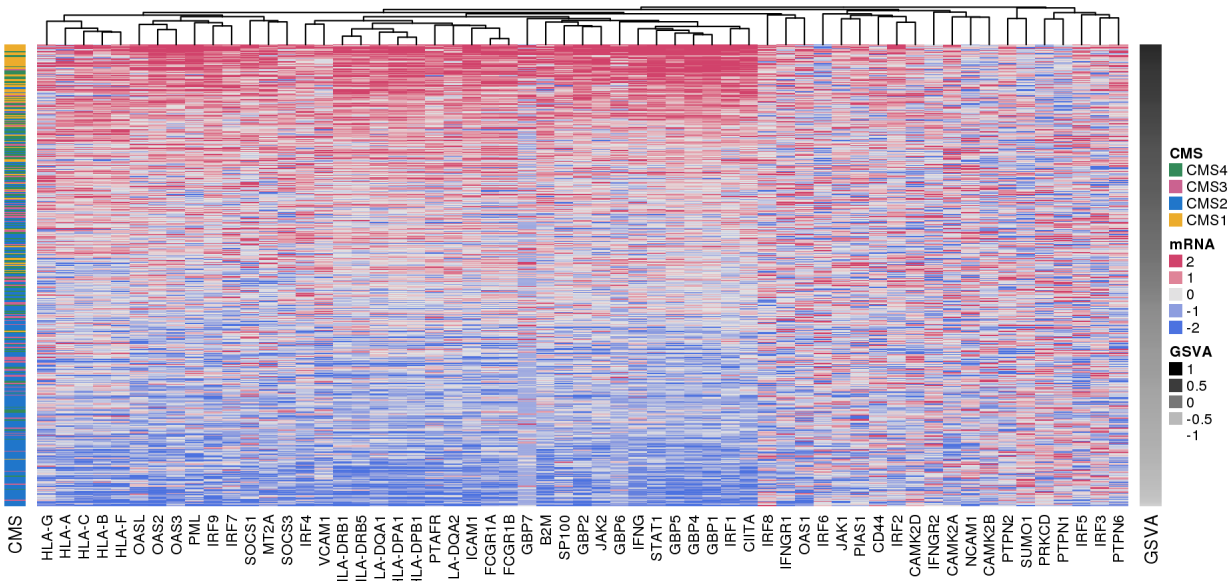


Figure 3.9: Heatmap of gene expression and GSVA score with samples sorted by GSVA score (REACTOME_INTERFERON_GAMMA_SIGNALING)

Totally 8 predictors are selected with table showing the selection probability, magnitudes, and marginal correlation with the pathway score listed in Figure 3.10. Regarding with the predictor names, "M_" represents the methylation factors. Thus, there are 7 methylation sites that are selected as the switches and 1 microRNA. Among all 8 predictors, "M_HLA.DQA2_cg11706729" has the highest selection probability and is negatively correlated with the pathway score. If we further check if this predictor significantly differentiates the CMS groups, we compute its corresponding anova p value as smaller than 0.05 with violin plot shown in Figure 3.11. Moreover,

we checked the model performance by measuring its R^2 and concordance correlation coefficient with scatter plot showing the exact pathway score values as well as the predicted values as shown in Figure 3.12. The high R^2 and CV_CCC values indicate that our model predicts well and yield to great parsimonious given the sparse predictor set.

Interpretation: Study shows that Interferon-signaling pathway has important impact on colon and rectal cancer types in their risk and survival due to its genetic variation [100]. Methylation site cg05141234 is selected as upstream factor mapped with genes HLA.DPA1 and HLA.DPB1. HLA gene family functions in forming group of related proteins named human leukocyte antigen (HLA) complex that helps the immune system in the body. It has been reported that the member gene HLA.DPA1, with the changes in gene expression, is likely to be one of the representatives in causing colon adenocarcinomas [97]. Methylation site that is associated with gene OAS3 is also selected to be important. Study has demonstrated that OAS with full name oligoadenylate synthetase that regulates ribonuclease L and PKR (protein kinase R) and with the regulated RNaseL and PKR being found to be among several IFN-induced genes that attribute direct in vivo antiviral activity [95]. Two methylation CpG sites related with gene STAT1 significantly and negatively influences the pathway activity. Study shows that the STAT1 gene which belongs to Signal Transducers and Activators of Transcription (STATs) family has been found to have decreased expression levels in transformed intestinal epithelial cells in colorectal cancer [57]. Moreover, this finding consists with STAT1's tumor suppressor nature. The microRNA miR.1296 has been recently discovered to serve as tumor suppressor role in a subtype of breast cancer [91], and also found to be prognostic for distant metastasis-free survival (DMFS) in T2-T3N0 colon cancer [15].

upstream factors	selection_probability	OLS_coefficient	correlation
M_HLA.DPA1_HLA.DPB1_cg05141234	0.522	0.1137	0.607
M_HLA.DQA1_cg19136673	0.521	-0.0531	-0.7146
M_HLA.DQA2_cg11706729	0.955	-0.0714	-0.6737
M_ICAM1_cg10651168	0.553	-0.0441	-0.3734
M_OAS3_cg19371652	0.639	-0.0547	-0.6439
M_STAT1_cg00676801	0.505	-0.0475	-0.5828
M_STAT1_cg03110996	0.83	-0.0253	-0.6011
hsa.miR.1296	0.776	-0.0803	-0.31

Figure 3.10: Selected upstream factors and their selection probability, coefficient from OLS model and marginal pearson correlation for pathway REACTOME_INTERFERON_GAMMA_SIGNALING

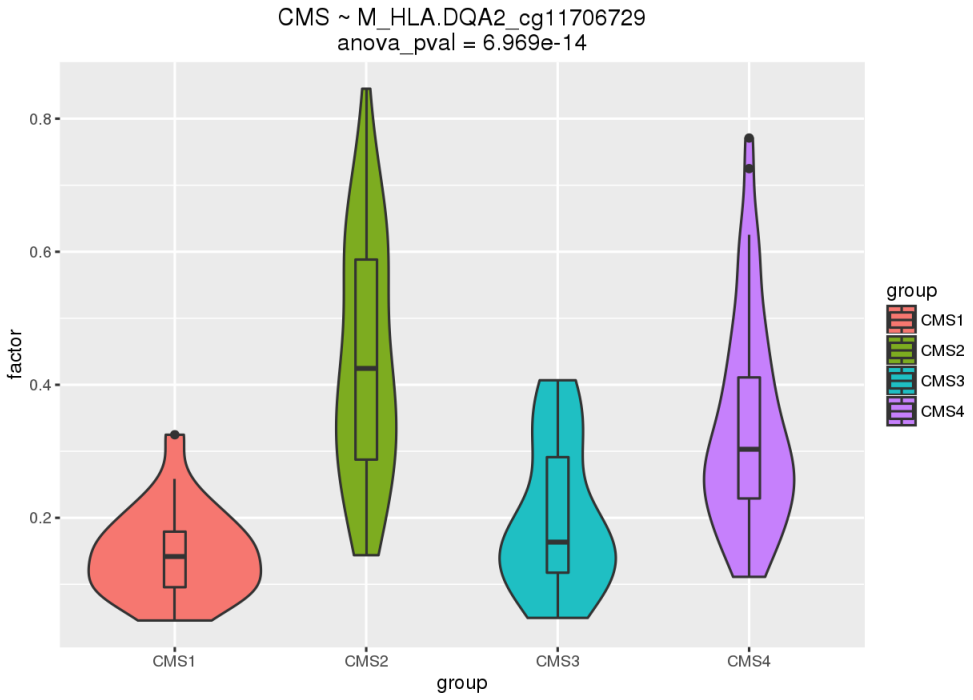


Figure 3.11: Violin plot of the methylation HLA.DQA2_cg11706729 across CMS groups

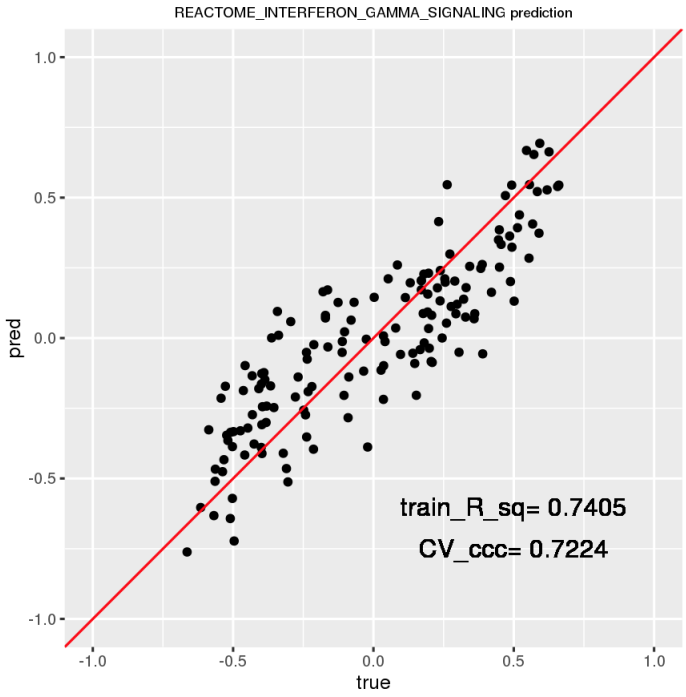


Figure 3.12: Scatter plot of the prediction performance for pathDrive modeling (REACTOME_INTERFERON_GAMMA_SIGNALING)

3.5 Shiny App Application

We established an interactive web app "Shiny App" to illustrate the results, simultaneously, this App provides people a way to retrieve their interested pathways, the general pathway properties, as well as specific pathDrive modeling results. The App could be accessed under `http://qcprpudev6:3833/pathDrive_YZhang/`, and for Colorectal Cancer, people can target different pathway collections, their properties in differentiating CMS groups. For each pathway, the following list shows the items that could be retrieved.

- Membership genes that belong to the pathway
- Heatmap showing mRNA and GSVA scores labeled with CMS groups, the samples are sorted based on GSVA scores.

- Boxplot showing GSVA score for different CMS groups for 3 data sets (Affy, TCGA and Integromics)
- Scatter plot showing predictive accuracy with R^2 within training set and 5-fold cross validation concordance correlation coefficient.
- Table of the selected upstream factors, their selection probability, coefficient and marginal correlation with GSVA score.
- For each selected upstream factor, it can lead to its violin plot for different CMS groups with ANOVA test being implemented. And the pathways that select this factor are also listed in a table.

3.6 Discussion

Our pathDrive framework provides novel biological formulation in identifying a sparse set of regulatory factors that influence pathway activity that is measured based on gene expression levels. By detecting the key genetic and epigenetic switches that may be the potential drivers of the functional activity, our proposed framework has several key advantages: (1) our model involves pathway level measurement as response rather than individual gene level summaries, which describes higher level representative of the biological process; (2) the formulation accounts for the natural information flow of DNA $\rightarrow$ mRNA $\rightarrow$ protein regulated by genetic and epigenetic factors; (3) the framework incorporates latest multiplatform genomic factors as potential regulators; (4) the modeling part utilizes efficient and interpretable feature selection algorithm. Although we mainly applied our methodology into the case study in colorectal cancer types with 4 subtypes, our method framework is general and can be directly applied into other disease domain with pathway score generated and calculated from the corresponding gene expression and upstream factor matrix

aggregated from their multiplatform genomic data profiles.

We applied our methodology framework into the analysis CRC with 4 CMS groups incorporated into the study. Important pathways that significantly differentiate the CMS subgroups were selected, and for each, we model the relationship between pathway score and the genetic and epigenetic factors. Result shows that parsimonious could be achieved in that sparse set of the predictors were selected that can explain large variation of the pathway score for most of the pathways. Moreover, results for some specific targeted pathways that were also illustrated, such as pathway Hallmark INF-α Response, Reactome INF-α/β signaling, Reactome INF signaling and etc. For example, for Reactome INF-γ signaling pathway, factors M_HLA.DQA2_cg11706729, M_STAT1_cg03110996, M_STAT1_cg00676801, miR.1296 and etc. with some of the factors selected and visualized and based on literature review, M_STAT1_cg03110996 lies in CpG shore within promoter region, it functions as active enhancer. It has been demonstrated that STAT1 has prosurvival effects on KRAS colon cancer cells by regulating pathways that initiate mRNA translation. Other important pathways for CRC were summarized with interactive Shiny App as discussed in Section 3.4.3.

To summarize our study, we establish a general pathDrive modeling framework, outline the pathway score calculation, upstream factor formulation, modeling strategy as well as results interpretation. We applied our framework to identify potential CMS-specific therapeutic targets, obtained the results for overall targeted pathways and for some specific pathways. Our findings offer potential drug discovery biomarker targets and can be further investigated in the future biological experiment and analysis. Our Shiny Application provides a useful tool for biologists/bioinformaticians in our result searching with more user-defined and convenient interface, which facilitate further investigation and research in this area.

We admit that our analyses yield to multiple limitations. From data preprocessing perspective, the selection of miRNA to include in the model was implemented using an unsupervised approach based on targetHub [69] rather than a supervised approach checking directly at mRNA-miRNA correlation in our data sets. From modeling perspective, we only presented results using the LASSO for feature selection. As mentioned in Section 3.2.3, technically, we can apply any well-established feature selection algorithms, however, here in this thesis chapter, we do not include the implementation of model comparison based on our real data sets. Implementing different models and doing comparison evaluated by multiple predictive criteria will be left to future work.

Several possible future directions could be explored based on our proposed framework. One direction could be including the complicated interaction terms between different platforms or within each platform into the model, such as interaction between neighboring methylation CpG sites, microRNA and etc. Another possible advancement may be incorporating nonlinearity in the relationship modeling which allows more flexibility. One other direction may be exploring more statistical models and summarizing the selection and prediction results with ensembling methods. We leave these tasks for future work.

Chapter 4

BLINK: Bayesian Linked Regression Models for Pan-Cancer Genomic Data Integration

4.1 Introduction

Rapid development in genomic technologies and worldwide collaboration in generating comprehensive and diverse genome-wide database, such as The Cancer Genome Atlas (TCGA) [108], motivate integrative analysis of multi-dimensional molecular profiles within and across tumor types. These analyses aim to delineate hidden associations among different genomic platforms, and with clinical characteristics. Besides the high throughput genomic data sets available, more comprehensive study eager to involve the analysis of protein expression data. Given that protein levels, which contain complex genomic and transcriptomic aberrations with translational and post-translational regulation, can hardly be investigated merely through the analysis of genomic and transcriptomic, the reverse-phase protein arrays (RPPA), a quantitative antibody-based technology, has been applied in functional protein

analysis. With this advanced technology, people have brought in a high-throughput proteomics database of the samples from TCGA, which is publically available through the Cancer Proteomics Atlas (TCPA) [66].

Given the availability of high-quality genomic and proteomic data sets, more comprehensive analysis including correlation studies between protein levels and other genomic related profiles, have been implemented: e.g, Akbani *et.al* [4] shows relatively high correlation between protein and mRNA on HER2; Zhang *et.al* [121] discovered that protein level, measured through PI3K/AKT/mTOR pathway activity, was found to be associated with mutations or copy alteration. Thus, our initial interest lies in finding upstream transcriptomic and genomic drivers that affect proteomics, so as to elucidate underlying biological protein regulatory mechanisms for a specific cancer type. Furthermore, study shows that analyzing proteomic processes with upstream transcriptional factors across different tumour types is of great necessity compared with disease-specific studies which may neglect potential commonalities and differences in molecular mechanisms. And recent studies demonstrate the existence of this commonality, e.g, Milewska *et. al* [74] found that 12% of their cancers have ERBB family somatic mutations based on the analysis of more than 14,000 patients. Thus, our further interest lies in detecting the upstream transcriptomic and genomic factors that drive proteomic processes for different cancer types. However, effective signal detection problem lies in that rare cancer types do not have enough sample size, while common cancers, with comparatively larger sample sizes, tend to have stronger power for signal detection. With this unbalanced sample size issue and the existing commonality discovered in genomic drivers, in this Chapter, we target to detect genomic upstream proteomic drivers for multiple cancer types, borrowing strength of the signals from common cancers to rare cancers based on their hidden commonality/similarity of the regulators, that improves the sensitivity and specificity

of the signal detection, and at the same time, promotes the learning of the similarity in molecular mechanisms across different cancer types.

Given biological problem as described above, statistically, our goal is to characterize key covariates that influence the responses for multiple strata while borrowing strength from the similarity structure inferred by the common covariates that are shared across strata. The regression stratum can be of different forms, such as gender, nationality, race and etc. In our case study, we take cancer type as the stratum. Classic approach for allowing covariates to vary across strata is through the formulation of a large single regression with the introduction of interaction terms among multiple strata [3]. However, this approach does not give a global sense of the similarity in the selected covariates among strata, and the interaction lies in coefficient estimates rather than in covariate selection. Another common way people address the problem is to apply separate model inference procedure for each stratum. However, for rare cancer, this approach will consequently lead to low sensitivity or specificity due to lacking of enough sample size. And the approach neglects the study of potential similarity among covariates across the strata. When considering borrowing strength, alternative common way is to implement the inference with large pooled data that integrates all strata samples. However, this approach results in high false positive rate due to overly borrowing strength especially for strata that are not similar in their true signal structure. Therefore, we mainly consider adaptively borrowing strength in covariate selection across multiple strata through some sort of "correlation" which can be implemented in different ways. One could utilize Seemingly Unrelated Regression (SUR) [120] model where they improve parameter estimation efficiency by borrowing strength through contemporaneous cross-equation error correlation and later has been extended to Sparse seemingly unrelated regression (SSUR) [111] that can be applied into high dimensional cases. However, both of them are limited in direct

borrowing strength for signal detection and exploring its selection similarity. Another approach that can be applied is through the correlation in magnitudes. That is, basically, when stacking the multiple strata equations on top of each other as shown in Equation (1), the entire system can be regarded as a large single regression problem, to borrow strength, people can apply different grouped shrinkage methods such as grouped lasso [119], normal gamma shrinkage prior [32], horseshoe prior [19] and etc. for possible related covariates. However, such an approach can be limited in allowing the flexibility for the stratum-specific coefficients to vary in signs and magnitudes.

In this Chapter, we developed novel Bayesian linked regression modeling framework to the problem of variable selection for multiple strata. The linkage lies in the probabilities of covariate indicators, which is learnt via a Markov Random Field (MRF) prior [65]. This specific prior encourages the selection of a covariate for the regression given the same covariate is detected in related regressions. Unlike the aforementioned approaches, the regression relatedness is imposed and learnt from signal detection aspect, more specifically, higher relatedness or similarity indicates more common selected covariates. This pairwise relatedness is learnt and modeled by putting a spike-and-slab prior on parameters that represent the similarity measurements, with their inclusion posterior probabilities indicating the corresponding pairwise strata similarity. To detect stratum-specific signals, we borrow strength of the relatedness by incorporating the similarity parameters into MRF prior setting, allowing adaptively sharing information between strata. Moreover, we also formulate variable-specific informative inclusion prior probability that encourages proper sparsity. Our proposed framework allows adaptively borrowing strength across related strata and infering their relatedness from common signals. In our simulation study, we demonstrate the advantage of our model in that it leads to better variable selection performance compared with other alternative approaches,

and the capability of learning the similarities. We find higher accuracy in variable selection especially for strata with relatively small sample sizes and with high hidden similarity linkage with other strata regressions. Finally, to illustrate our methodology, we implement an integrative grouped pan-cancer analysis in detecting driving genomic and transcriptional upstream factors that characterize RPPA based proteomic pathway activities, and simultaneously inferring which cancers in our group share more common proteomic regulative drivers for a specific pathway.

The rest of the thesis Chapter is organized as follows. We introduce our Bayesian linked regression model framework in Section 4.2 including formulation, similarity structure and prior settings for the main parameters and hyperparameters. Posterior sampling and Model selection will be presented in Section 4.3. The simulation results are demonstrated in Section 4.4 and the application of real pan-cancer integrative analysis is described in Section 4.5. We conclude this thesis Chapter with brief discussion in Section 4.6.

4.2 Statistical Model and Methods

Figure 4.1 illustrates the general flow of our modeling framework. Our joint model mainly contains two levels: Regression Level and MRF Process Level. The first level involves linear regression models with specific prior settings for the corresponding coefficients. The second level denotes the similarity parameters that link the models defined in the first level, and establishes the prior settings for the similarity parameter matrix that could be inferred from the joint model. The specific model settings, parameter definitions, prior and posterior distributions, and the model selection components are described in the following subsections.

4.2.1 Model at Regression Level

Our goal is to analyze key covariates that affect the dependent variable across multiple strata. More explicitly, suppose we construct a linear regression problem with K different strata. Thus, considering a single observation and a stratum, the model can be expressed using the form

$$y_{ki} = \sum_{j=1}^{j=P} x_{kij}\beta_{kj} + \epsilon_{ki}, k = 1, 2, ..., K; i = 1, 2, ..., n_k; j = 1, 2, ..., P \qquad (4.1)$$

where y_{ki} denotes the i^{th} observation for the k^{th} stratum, which is to be linearly explained by the covariates with the total number represented as P. Our methodology framework deals with the cases where we consider same regressors, in their variable types, but not their actual values. However, we allow different sample sizes for different cancer types. x_{kij} is the j^{th} predictor of the i^{th} observation for the k^{th} stratum, the corresponding coefficient is denoted as β_{kj}. ϵ_{ki} is the random error component of the i^{th} obervation for k^{th}. K is the total number of the strata. n_k denotes the sample size for the k^{th} stratum.

The K strata can be separately expressed as the above equation, forming totally K equations. However, to be more explicit, the K equations could be compactly expressed as

$$\begin{bmatrix} \boldsymbol{y_1} \\ \boldsymbol{y_2} \\ \boldsymbol{y_3} \\ \vdots \\ \boldsymbol{y_K} \end{bmatrix} = \begin{bmatrix} X_1 & & & \dots & \\ & X_2 & & \dots & \\ & & X_3 & \dots & \\ \dots & \dots & \dots & \dots & \dots \\ & & & \dots & X_K \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta_1} \\ \boldsymbol{\beta_2} \\ \boldsymbol{\beta_3} \\ \vdots \\ \boldsymbol{\beta_K} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\epsilon_1} \\ \boldsymbol{\epsilon_2} \\ \boldsymbol{\epsilon_3} \\ \vdots \\ \boldsymbol{\epsilon_K} \end{bmatrix} \qquad (4.2)$$

To estimate the coefficient vector $\boldsymbol{\beta} = \{\boldsymbol{\beta_1}, \boldsymbol{\beta_2}, ..., \boldsymbol{\beta_K}\}$, different mechanisms to borrow strength across multiple regressions could be applied. One can apply SUR model by incorporating the inter-correlation in error vectors $\boldsymbol{\epsilon} = \{\boldsymbol{\epsilon_1}, \boldsymbol{\epsilon_2}, ..., \boldsymbol{\epsilon_K}\}$, which leads to comparatively higher parameter estimating efficiency. However, this approach does not directly borrow strength from variable selection aspect and has the limitation of detecting the common variables that are selected by different strata regressions. Another common way is to borrow strength via the correlation of the magnitudes, which limits the flexibility in their signs and magnitudes. However, our model borrows strength via variable selection indicators across the regressions while allowing flexibility in coefficients, and adaptively detects the similarity of the selected features.

In our case study, X_k represents the covariate matrix with dimension $n_k \times P$ where n_k denotes the sample size for cancer type k, and P is the total number of the predictors that are included all across the cancer types. $\boldsymbol{y_k}$ represents RPPA based pathway activity score for cancer k, and X_k comprises the corresponding mRNA, methylation, copy number and microRNA factors matching with the genes belonging to that pathway. Our main goal is to select key molecular upstream factors that affect RPPA based pathway activity while borrowing strength across K cancer types and assessing the similarity in their signal selection simultaneously. More details of the application are described in Section 4.5.

4.2.2 Linking Regression via Markov Random Field Process

Considering a simple regression model for each stratum as described in Equation (1), our key target is to detect the significant covariates as mentioned in the introduction. To implement variable selection, we adopted the Bayesian "spike and slab" approach [75] [29] which defines a binary latent variable $\gamma_{kj} \in \{0, 1\}$ for each coefficient param-

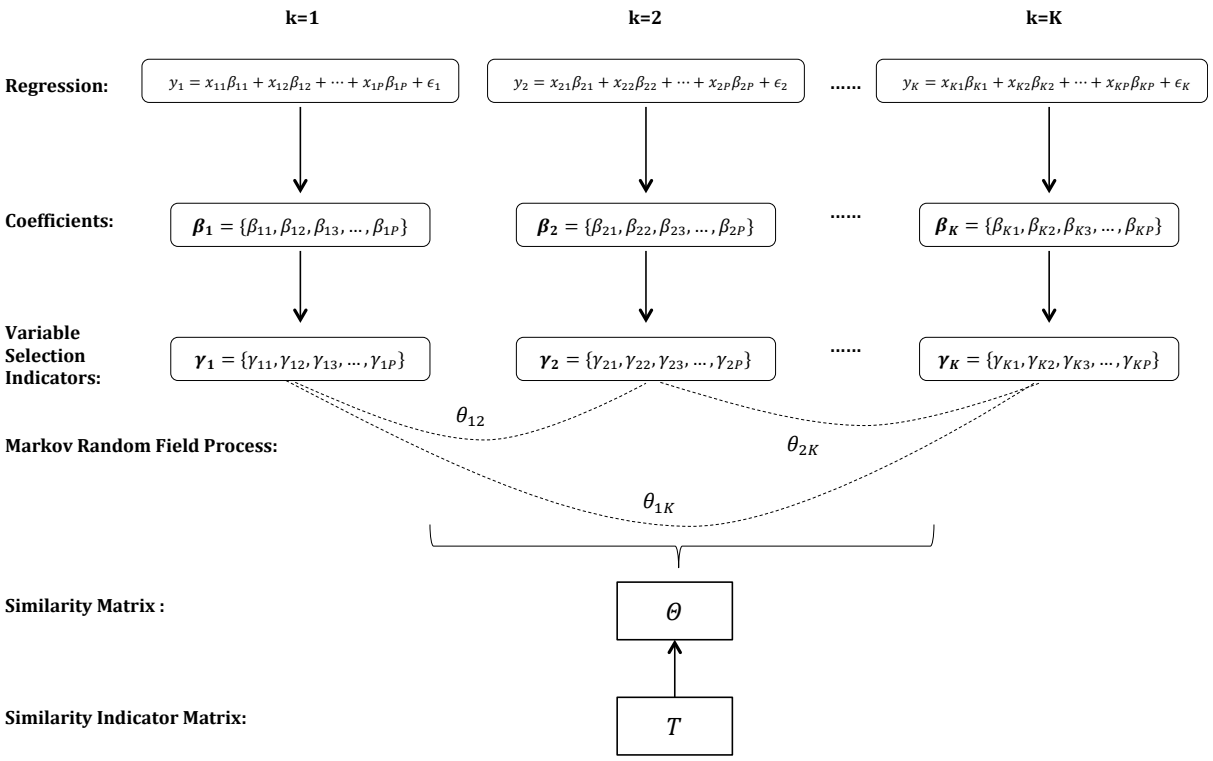


Figure 4.1: General modeling flow: Models at Regression level with K strata, each with coefficient vectors which are inferred by variable selection indicators; Markov Random Field Process incorporates the linkage of the indicators through similarity matrix denoted as Θ, with each elements inferred from similarity indicator Matrix.

eter β_{kj} as illustrated in Figure 4.1. In this particular framework, $\boldsymbol{\beta_k}$ depends on $\boldsymbol{\gamma_k}$ with each element independently following conjugate Gaussian mixture priors:

$$\beta_{kj} \sim (1 - \gamma_{kj})\delta_0 + \gamma_{kj}\mathcal{N}(0, c\sigma_k^2) \tag{4.3}$$

where δ_0 represents a point mass at zero. Residual variance σ_k^2 follows Inverse Gamma prior which is conjugate:

$$\sigma_k^2|\gamma_k \sim InverseGamma(e/2, ef/2) \tag{4.4}$$

and c represents a scalar that could be adjusted based on the magnitude and the sparsity of the signals. γ_{kj} follows a certain distribution that will be mainly discussed in the following sections. With the prior distribution, the entire variable selection procedure is then implemented based on stochastic search over the marginal posterior space where $\boldsymbol{\gamma_k}|\boldsymbol{Y} \in \{0, 1\}^P$.

In stead of individually assigning Bernoulli prior for each selection indicator vector $\boldsymbol{\gamma_k}$, in our approach, we link the multiple strata regressions by introducing dependence among each $\boldsymbol{\gamma_j}$, which takes the vertical direction of $\boldsymbol{\gamma_k}$ across K regressions. More specifically, $\boldsymbol{\gamma_j}$ denotes the selection indicator vector across K strata for the j^{th} predictor. To impose the selection dependence/linkage for each predictor, in our approach, we set up MRF prior on through the application of MRF prior settings on $\boldsymbol{\gamma_j}$. The particular prior was originally established and applied in Bayesian variable selection in linear regression context [65] based on the assumption that the covariates lie in an undirected graph, and with the hidden structure of the covariate matrix incorporated into the framework. The MRF prior that we developed acts similarly wherein we incorporate the structure for each variable across multiple regressions allowing similar strata select similar covariates. Moreover, the MRF prior was

applied and implemented in linking multiple graphs by Peterson et al., 2015 [90] and has been demonstrated having proper properties.

Specifically, consider one predictor j, the prior probability that we assign to the indicator vector $\boldsymbol{\gamma_j}$ can be expressed as:

$$p(\boldsymbol{\gamma_j}|\nu_j, \Theta) = C(\nu_j, \Theta)^{-1} exp(\nu_j \mathbf{1}^T \boldsymbol{\gamma_j} + \boldsymbol{\gamma_j}^T \Theta \boldsymbol{\gamma_j}) \tag{4.5}$$

where $1 \leq j <= P$ denoted as the j^{th} predictor, $\boldsymbol{\gamma_j}$ denotes the latent inclusion indicator of the j^{th} variable across K strata regressions. $\mathbf{1}$ represents the unit vector with dimension K, ν_j represents the same variable specific parameter across K, $\boldsymbol{\Theta}$ denotes the symmetric matrix with dimension $K \times K$ representing the similarity among covariate selection across K regressions. More specifically, matrix $\boldsymbol{\Theta}$ is off-diagonal matrix with each element representing the pairwise similarity as magnitude in terms of the variable selection of the corresponding regressions.

More specifically, based on the MRF prior, the normalizing constant is denoted as $C(\nu_j, \Theta)$ where

$$C(\nu_j, \Theta) = \sum_{\boldsymbol{\gamma_j} \in \{0,1\}^K} exp(\nu_j \mathbf{1}^T \boldsymbol{\gamma_j} + \boldsymbol{\gamma_j}^T \Theta \boldsymbol{\gamma_j}) \tag{4.6}$$

If there for all strata regressions, no features are selected, the prior probability becomes

$$p(\boldsymbol{\gamma_j} = \mathbf{0}|\nu_j, \boldsymbol{\Theta}) = \frac{1}{C(\nu_j, \Theta)}$$

Specifically, given the prior, the inclusion probability of the j^{th} predictor for the k^{th} regression given the inclusion of the j^{th} predictor in the remaining regressions is

$$p(\gamma_{kj}|\{\gamma_{mj}\}_{m\neq k}, \nu_j, \Theta) = \frac{exp(\gamma_{kj}(\nu_j + 2\sum_{k\neq m}\theta_{km}\gamma_{mj}))}{1 + exp(\nu_j + 2\sum_{m\neq k}\theta_{km}\gamma_{mj})} \tag{4.7}$$

4.2.3 Variable Selection Structure Similarity

$\boldsymbol{\Theta}$ denotes the selection similarity matrix, representing how similar two regressions perform in terms of selecting the same variables, if considering two strata, more variables that are selected by both of them leads to higher similarity. The elements in this nonzero off-diagonal matrix represent the magnitude of the pairwise similarity. In our analysis, the similarity matrix is inferred from data. Considering the fact that two strata may have similar driving predictors at different levels, or not having any common ones, the similarity matrix needs to be flexible enough to handle the exisitence of the similarity and the magnitude. Driven by this motivation, we place a spike-and-slab prior on each element of $\boldsymbol{\Theta}$, denoted as θ_{km}, by introducing a corresponding latent indicator matrix $\boldsymbol{T}$, with each element denoted as τ_{km}, as illustrated in Figure 4.1. To effectively represent the similarity, θ_{km} needs to be either zero, or positive, to induce the better discrimination of zero and positive values, we put $Gamma(\alpha_\theta, \beta_\theta)$ distribution when $\alpha_\theta > 1$ as the 'slab' portion of the spike-and-slab. Johnson et al., 2010 [51] described the performance in model selection when putting nonlocal prior in details.

The latent variable τ_{km} represents if regression k and regression m are related in terms of selecting common features. The mixture prior for element θ_{km} can be

written as

$$p(\theta_{km}|\tau_{km}) = (1-\tau_{km})\delta_0 + \tau_{km}\frac{\beta_\theta^{\alpha_\theta}}{\Gamma(\alpha_\theta)}\theta_{km}^{\alpha_\theta-1}e^{-\beta_\theta\theta_{km}} \tag{4.8}$$

where α_θ and β_θ are fixed hyperparameters. The binary latent variable τ_{km} is given independent Bernoulli prior distribution

$$p(\tau_{km}|w) = w^{\tau_{km}}(1-w)^{1-\tau_{km}} \tag{4.9}$$

where w denotes a fixed hyperparameter with value lies in [0,1]. Further, the entire similarity matrix $\mathbf{\Theta}$ follows the distribution

$$p(\mathbf{\Theta}|\boldsymbol{\tau}) = \prod_{k<m} p(\theta_{km}|\tau_{km}) \tag{4.10}$$

4.2.4 Marginal Variable Selection Prior

The equation (1) shows that for a given k, the prior probability of the inclusion of predictor j given that the predictor is not selected by any rest of the regressions, or $\theta_{km} = 0$ for all $m \neq k$ is

$$p(\gamma_{kj}|\nu_j) = \frac{e^{\nu_j}}{1+e^{\nu_j}} \tag{4.11}$$

where ν_j serves as the parameter that adjusts the sparsity of all the regressions and also allows the flexibility in selection of particular predictors. Smaller ν_j leads to lower marginal inclusion probability of the j^{th} predictor for all regressions, so as to impose the overall sparsity, which in application, will give us more interpretable results. We denote the above marginal probability as q_j and to induce the flexibility, we set up a prior for q_j in our analysis. Since the prior needs to be flexible enough to cover the probability range (0,1), we chose Beta prior distribution $Beta(a,b)$. As can be indicated, parameter $\nu_j = logit(q_j)$, the prior for q_j also serves the prior for ν_j with

the transformed prior distribution as

$$p(\nu_j) = \frac{1}{Beta(a,b)} \frac{e^{a\nu_j}}{(1+e^{\nu_j})^{a+b}} \tag{4.12}$$

where a and b are fixed hyperparameters that adjust the prior value of q_j. To incorporate the prior belief of the sparsity of the overall regressions, a prior which leads to lower q_j is favored, such as $Beta(1,7)$. In application, the specific values can be adjusted taking account the sparsity of the specific case. If it is believed that most of the predictors are not selected by any of the regressions and the ones that are selected are more likely to be the shared ones, people may consider set up a prior that favors small value of q_j with the prior that leads to larger θ_{km} values.

4.3 Posterior Sampling and Model Selection

Considering we have K regressions, for each regression, we have P predictors. Given the above prior settings and the likelihood, the full distribution can be written as

$$\prod_{k=1}^{K} \Big\{ p(\boldsymbol{Y_k}|\boldsymbol{\beta_k}, \sigma_k^2) p(\boldsymbol{\beta_k}|\boldsymbol{\gamma_k}, \sigma_k^2) \prod_{j=1}^{P} [p(\gamma_{kj}|\nu_j, \boldsymbol{\Theta}) p(\nu_j|a,b)] \Big\} p(\boldsymbol{\Theta}|\boldsymbol{\tau}) p(\boldsymbol{\tau}) \tag{4.13}$$

Where the posterior samples of our interested parameters can be acquired by constructing a MCMC sampling scheme.

4.3.1 Sampling Scheme

The entire modeling structure shown in equation (1) comprises K regressions, in our sampling scheme, we run the regression univariately, but taking account of the shared information incorporating similarity matrix $\boldsymbol{\Theta}$ into the prior setting of $\boldsymbol{\gamma}$.

As can be indicated, the top level of the MCMC scheme involves spike-and-slab for

each of the regression. Specifically, for the k^{th} regression, sampling $\boldsymbol{\beta_k}$ and $\boldsymbol{\gamma_k}$ based on proper prior information. Here, for the normal 'slab' portion, in our analysis, we applied independent prior setting as shown in equation (2), assuming each regressor effect follows the same distribution and is independent. Other widely used conjugate priors (g-prior and fractional prior) can also be used under some specific cases. To get more efficient and fast sampling of the coefficients of each regression, we adapt SSVS approach of George and McCulloch (1993) (cite). We then sample the similarity matrix $\boldsymbol{\Theta}$ and the indicator matrix $\boldsymbol{\tau}$ from the conditional posterior distribution applying Metropolis-Hastings method incorporating between-model and within-model moves, in another word, reversible jump MCMC sampling approach. Last step involves the sampling of $\boldsymbol{\nu}$ with traditional Metropolis-Hastings steps.

In summary, our MCMC sampling scheme can be described as follows with the detailed sampling procedure described in Appendix B.1. For each iteration t,

- Update $\boldsymbol{\beta_k}$, $\boldsymbol{\gamma_k}$ and σ_k^2 for each regression, $k = 1, 2, ..., K$.
- Update each element θ_{km} and τ_{km} for similarity matrix $\boldsymbol{\Theta}$ and indicator similarity matrix $\boldsymbol{\tau}$, $1 \leq k < m \leq K$.
- Update predictor-specific marginal inclusion probability parameter ν_j where $j = 1, 2, 3, ..., P$ and P is total number of the predictors for each regression.

4.3.2 Model Selection

Our approach to carry out the signal selection is via the estimation of the posterior probability of the $\gamma_{kj} = 1$, indicating the inclusion of the j^{th} variable for the k^{th} stratum, with the proportion directly calculated from MCMC interaction after burn-in samples. We then propose variable selection procedure by setting up a threshold δ and regard the corresponding covariate $\boldsymbol{x_{kj}}$ to be significant if the posterior proba-

bility of $\gamma_{kj} = 1$ larger than δ. In our model selection, we choose δ equal to 0.5 based on the median rule set by Barbieri and Berger [9]. To infer the magnitudes, we refer to MCMC summary of β_{kj}, which is calculated from the marginal posterior mean, at each iteration, conditioning on the inclusion of the corresponding covariate ($\gamma_{kj} = 1$). Similarly, to detect the similarity in variable selection between the m^{th} and the k^{th} strata, we refer to the summary of indicator parameter τ_{km} from MCMC samples, we conclude that there exists significant similarity if the correponding posterior probability larger than 0.5. The magnitude of the similarity can also be calculated based on the conditional posterior mean from MCMC iterations.

Another parameter that we are interested in is ν_j which, with its inverse logit, indicates the marginal inclusion of the j^{th} covariate across all the strata. The summary of this parameter vector demonstrates the baseline covariate inclusion probability across strata without borrowing strength of the similarity.

4.4 Simulation

In our simulation, we apply our methodology framework with different scenarios that vary in true sparsity and the standardized coefficients or beta coefficients (absolute coefficient standard deviation). We check the performance of our model in terms of the parameter convergence and the estimation accuracy. Then, we compare the our approach with other alternative models, and illustrate how our model outperforms others in detecting the signals with higher accuracy and in learning the similarity structure among multiple strata groups.

4.4.1 Performance Checking and Parameter Estimation

In this simulation, we consider 5 different strata ($K = 5$) and 50 total predictors ($P = 50$), forming 5 groups, each expressed with one regression formulation with different sample sizes. Specifically, we draw random design matrix $\boldsymbol{X_k}$ from the distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{I_k})$, for $k = 1, 2, 3, 4, 5$ respectively with sample size $n_1 = 20, n_2 = 50, n_3 = 70, n_4 = 100, n_5 = 150$. Also, we consider two degrees of the sparsity, one with 5 true signals for all strata (sparsity 5/50), one with 10 true signals (sparsity 10/50). Moreover, for each sparsity level, we incorporate 4 different standardized coefficient levels 0.4, 0.5, 0.7 and 1, where they merely differ in magnitudes with common residual standard deviation equal to 0.3 which is approximately the calculated standard deviation of our real data. Thus, the true coefficients are sampled from normal distribution respectively centered around 0.4×0.3, 0.5×0.3,0.7×0.3,1×0.3 with low standard deviation (0.01) which ensures small variation in the magnitudes. The response vector for each stratum is sampled from the distribution $\mathcal{N}(\boldsymbol{X_k}\boldsymbol{\beta_k}, 0.3\boldsymbol{I_k})$. Intuitively, larger sparsity leads to fewer shared signals, and higher standardized coefficient level indicates stronger signal, yields to higher accuracy in the signal detection.

Here in our context, we only illustrate the performance of the first scenario, which has sparsity 5/50 with standardized coefficient 0.5. The performance of other scenarios will be shown in Supplementary materials.

We construct the true signal structure that all 5 strata have the same number of the signals as shown in the left panel of Figure 4.2, however, they differ in overlapping signals, in another word, they have different pairwise similarity in signal structures. The calculated proportion of the signals shared between strata can be

regarded as empirical estimate of $\boldsymbol{\Theta}$, which is:

$$shared\ signal\ proportion = \begin{bmatrix} 1 & 0.67 & 0 & 0.43 \\ & 0.67 & 0 & 0.43 \\ & & 0.11 & 0.43 \\ & & & 0 \end{bmatrix}$$

For prior settings in the simulation, we assign $\alpha_\theta = 2$ and $\beta_\theta = 4$ for $Gamma(\alpha_\theta, \beta_\theta)$ as the prior distribution for the slab part of the mixture prior defined in Equation (8), the hyperparameters in this Gamma distribution could be adjusted conditioning on $\alpha_\theta > 1$ which results in a nonlocal prior mentioned in Section 4.2.4. Simultaneously, hyperparameters that result in large θ_{km} also needs to be avoided, so as to prevent the exhibition of phase transition described in details in Li and Zhang et al., 2010 [65]. Therefore, we consider 0.5 as the mean in the similarity magnitude given the inclusion of the pairwise similarity, that favors strong prior belief that the strata share more signals, together with small q_j prior probability ($a = 1, b = 9$ in Equation (12)) that forces sparsity in that 10% of the predictors can be selected as signals as the baseline across all the strata for all predictors, is reasonable in our setting. However, higher q_j prior probability and lower similarity prior probability can be applied if one believes in more baseline signals, but fewer shared ones across the strata.

In the modeling implementation, we ran MCMC sampler with 30000 iterations, the result summary comes from 15000 iterations as burn-in samples and 15000 for inference. To check the convergence, Figure 4.3 shows the traceplots of the number of the detected signals for each strata regression, which illustrates the convergence. The figure shows proper mixing without strong trend in the estimation of the number of signals. Moreover, the posterior probability of inclusion (PPI)of j^{th} predictor for the k^{th} stratum can be calculated by $PPI(\gamma_{kj} = 1) = \frac{1}{T-S} \sum_{S+1}^{T} \boldsymbol{I}(\gamma_{kj} = 1)$ where

$T = 30000, S = 15000$. Right panel of the heatmap in Figure 4.2 shows the signal PPI of all 5 strata for the same scenario in our simulation. The estimated pattern indicates that our PPI estimation approximately matches with the true signal. To further assess the detection accuracy, we applied receiver operating characteristic (ROC) curve that can demonstrate the performance of variable selection when discrimination threshold is varied, with the Area Under the Curve (AUC) calculated, as illustrated in Figure 4.4 for each stratum. The AUC was 1.00 for stratum 5, which indicates perfect AUC can be reached when sample size is large enough in which case the variable selection is weakly affected by the prior setting, which, satisfies our desire to have strata which are not influenced too much by borrowing strength given enough sample size to support the detection power, thus can be effective in leading effective and better signal selection for other strata with small sample sizes borrowed strength from these "strong" strata. The AUC was 0.9867 for stratum 4, 0.9956 for stratum 3 and 0.9733 for stratum 2, which illustrate high accuracy performance. Stratum 2 achieved 0.9689 AUC value, which is lower given its limited sample size, but still high enough to obtain the correct signal detection.

To test the detection stability and reproducibility, we implemented our model for 50 replicates, with the averaged TPR, FPR and AUC calculated for each stratum with the standard deviation with displayed in parentheses in Table 4.1. The table clearly illustrates high AUC and low FPR in general, and with great stability due to small standard deviation. We noticed that for the first stratum having smallest sample size, the standard deviation of the performance assessment is comparatively larger, which is reasonable and within the normal range given the situation.

Besides the signal detection for each stratum, we check the marginal PPIs for similarity matrix $\mathbf{\Theta}$ with each off diagnal elemented estimated as $PPI(\theta_{km} = 1) =$

$\frac{1}{T-S}\sum_{S+1}^{T}\boldsymbol{I}(\tau_{km}=1)$ where $T=30000, S=15000$ and $1\leq k<m\leq K$ and $\boldsymbol{I}$ represents the indicator function. We noticed that the matrix estimation approximately match with the real shared linkage proportion in a way that higher value represents larger commonality pairwisely. The posterior mean indicates the extent of the similarity, which also provides our general information of how similar in terms of the signal selection between and among all the strata. We also calculated the standard deviation of the similarity matrix based on 50 replicates, the low deviation demonstrates high stability.

Thus, our methodology framework can obtain proper convergence and strong reproducibility with performance results in high signal detection accuracy and comparatively informative similarity estimation.

	TPR	FPR	AUC
k=1	0.476 (0.193)	0.0067 (0.012)	0.9632 (0.035)
k=2	0.86 (0.158)	0.0182 (0.024)	0.9841 (0.025)
k=3	0.968 (0.084)	0.0138 (0.018)	0.9982 (0.004)
k=4	0.952 (0.086)	0.0156 (0.015)	0.9972 (0.006)
k=5	0.992 (0.04)	0.0102 (0.016)	0.9997 (0.001)

Table 4.1: Simulation Results on 50 constructed data sets. Averaged true positive rate (TPR), false positive rate (FPR) and area under curve (AUC) with the corresponding standard deviation across the data sets.

$$sd(PPI(\boldsymbol{\Theta})) = \begin{bmatrix} 0.03 & 0.0.3 & 0.02 & 0.03 \\ & 0.05 & 0.04 & 0.04 \\ & & 0.05 & 0.04 \\ & & & 0.04 \end{bmatrix}$$

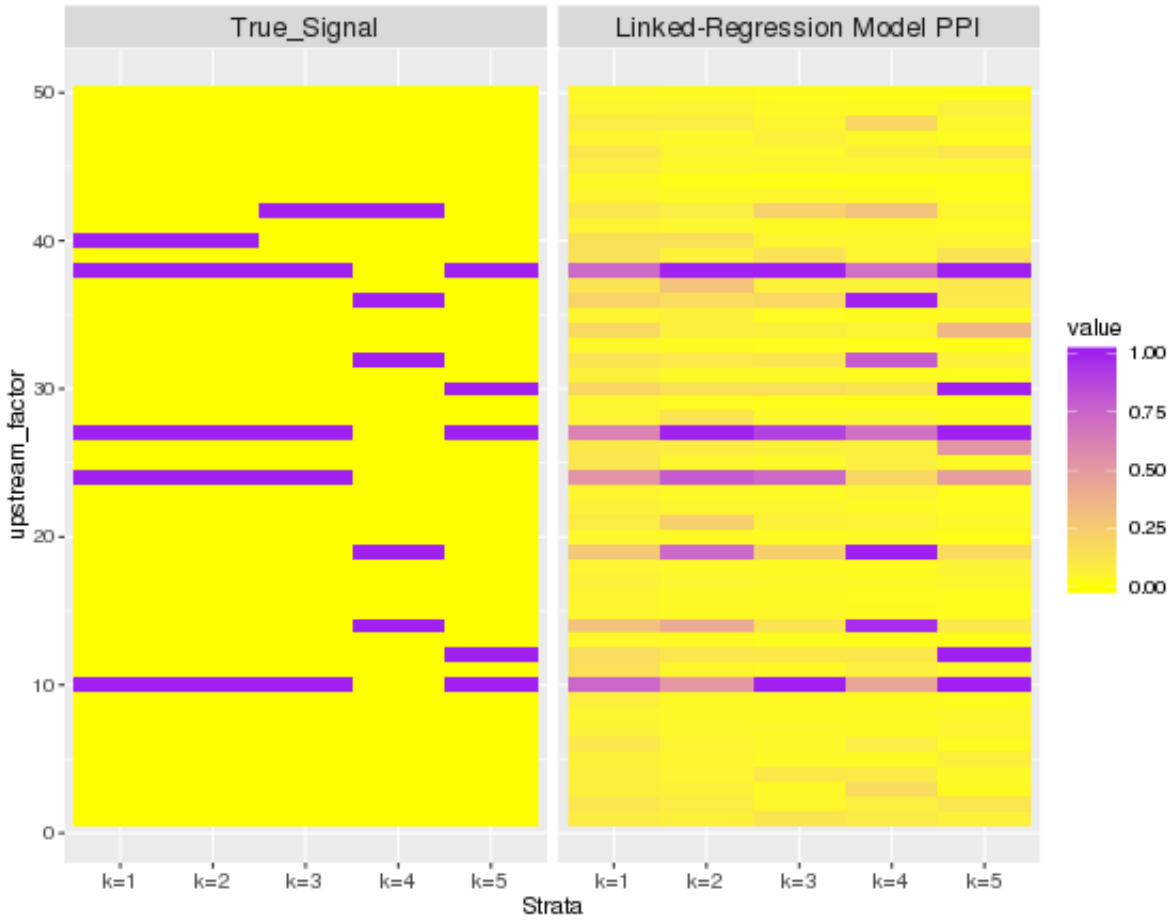


Figure 4.2: Modeling Result: left panel shows the true signal in our simulation setting with 50 predictors and 5 strata; right panel shows the posterior probability of inclusion (PPI) of our model, darker color indicated high posterior probability.

4.4.2 Performance Comparison

Further in the simulation, we implemented alternative approaches and compare their performances with ours. Specifically, we consider 4 other methods covering two types of possible solutions, one is separate regression estimation, the other one is pooled data with single regression. Two variable selection methods are considered for both of the solutions, spike and slab [75] and LASSO [107], which are typical methods respectively belong to Bayesian and Frequentist framework.

We generate 50 random normal datasets with the same similarity matrix $\mathbf{\Theta}$

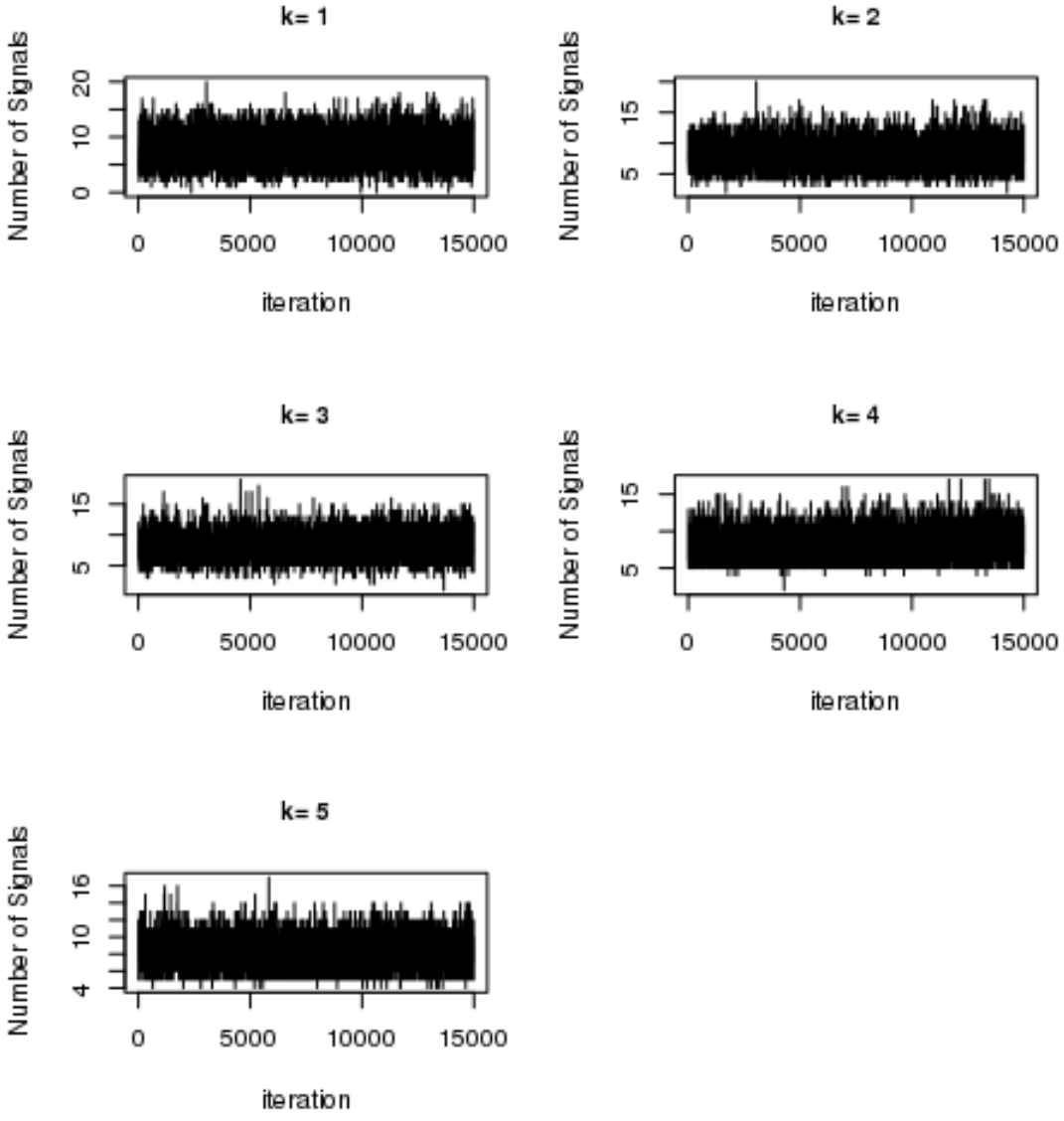


Figure 4.3: Traceplot: the traceplot of the number of the signals selected across MCMC samples with burn-in samples removed. It shows great convergence, which indicates proper mixing and the feasibility of our Linked-regression model.

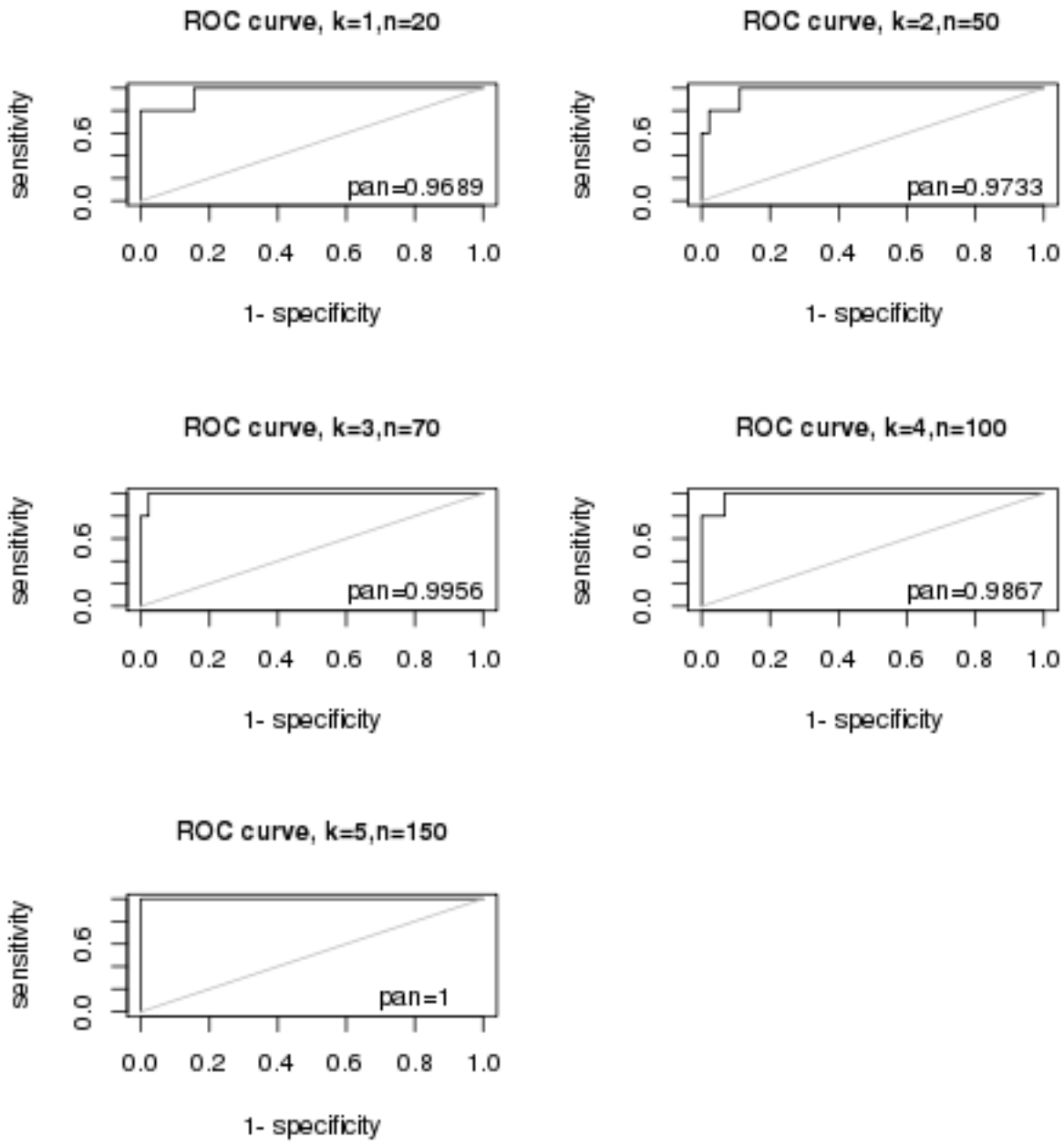


Figure 4.4: ROC curve: checking the ROC curve and the AUC of our model for the first senario, the computed AUC shows high accuracy of the signal detection for all 5 strata regression.

and implement the MCMC sampler and get the parameter summary based on MCMC samples for each replicate. The final results for the model comparison are summarized based on the analysis of 50 replicates . For each dataset, we apply our linked regression approach, marginal spike and slab method, marginal lasso method, pooled spike and slab method and pooled lasso method, we briefly name these methods in respectively in Table 4.2 as "BLINK", "mar_ss", "mar_lasso", "pool_ss" and "pool_lasso". For spike and slab, we set 0.5 as the prior variable inclusion probability and the result is estimated based on posterior probability. For lasso method, considering its randomness of the selection [107] we implement stability selection [73] so as to obtain the probability of the inclusion of the features. For all these methods, we implement with 30000 iterations and we regard the features to be significant if the posterior probability or the selection stability probability larger than 0.5.

Results on signal detection of different methods are shown in Table 4.1, which is based on standardized coefficient 0.5 at sparsity 5/50. The comparison of the performances for other scenarios is demonstrated in Supplementary materials. Figure 4.5 provides us visualized results that compare the AUC and TPR of different methods with the same FPR 0.05. It demonstrates that the averaged AUC and TPR of our model overperform the those of other alternative models for the first three strata when the sample size is relatively small. For last two strata with larger sample sizes, our linked-regression model tends to be slightly lower, but similar in general. The result meets our expectation that our model benefits more for the strata having smaller sample size with weaker strength in signal detection, for large sample strata, our model tends to make little influence on the posterial inference, so as to guarantee effective borrowing strength, in another word, we eager to borrow the information from the right signals detected for strong strata to weak strata.

To further check the comparison of the stability of the estimating results, Table 4.2 provides the averaged AUC, FPR and TPR with the corresponding standard deviation for all 5 models with fixed FPR controlled at 0.05. It shows that for strata with smaller sample size (eg. $k = 1$,$k = 2$ and $k = 3$), the TPR as well as the AUC calculated from our method uniformly outperforms that of others, leaving similar performance or slightly weaker performance for strata with large sample size. Our model is demonstrated to be stable and reproducible given that it has comparatively small standard deviation across 50 replicates than that of the other methods.

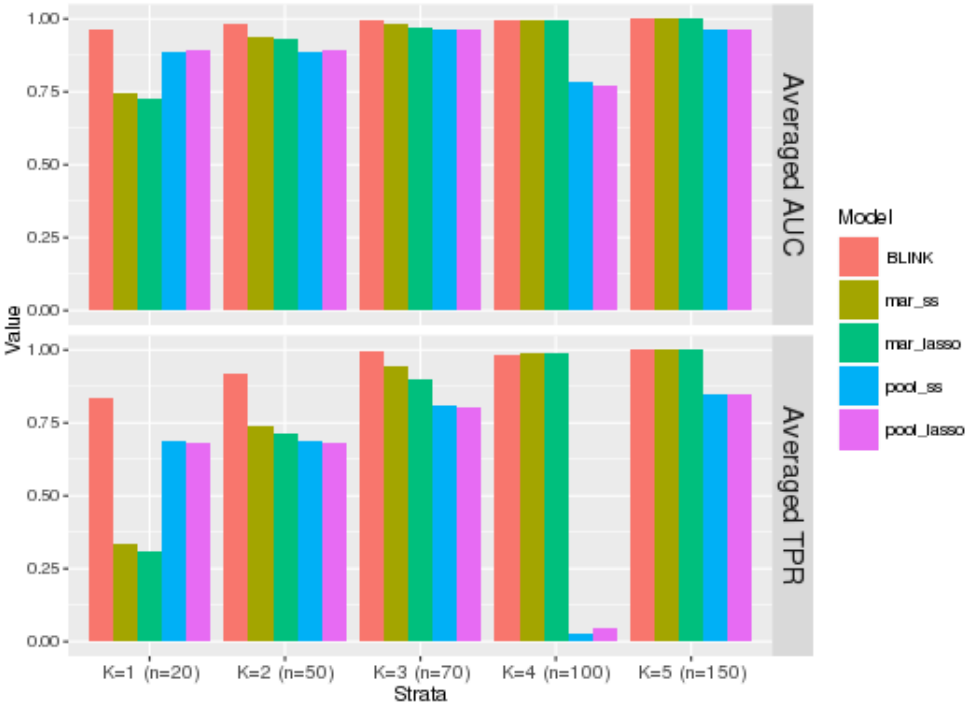


Figure 4.5: AUC and TPR Checking for model comparison with fixed FPR controlled at 0.05

	K=1 (n=20)			K=2 (n=50)			K=3 (n=70)		
Model	TPR (SE)	FPR (SE)	AUC (SE)	TPR (SE)	FPR (SE)	AUC (SE)	TPR (SE)	FPR (SE)	AUC (SE)
BLINK	0.832 (0.142)	0.05 (0)	0.9632 (0.035)	0.92 (0.128)	0.05 (0)	0.9841 (0.025)	0.996 (0.028)	0.05 (0)	0.9982 (0.004)
mar_ss	0.336 (0.254)	0.05 (0)	0.7469 (0.124)	0.736 (0.254)	0.05 (0)	0.9365 (0.07)	0.944 (0.107)	0.05 (0)	0.9836 (0.029)
mar_lasso	0.308 (0.233)	0.05 (0)	0.7244 (0.134)	0.716 (0.249)	0.05 (0)	0.9303 (0.075)	0.896 (0.141)	0.05 (0)	0.9716 (0.046)
pool_ss	0.688 (0.141)	0.05 (0)	0.8864 (0.069)	0.688 (0.141)	0.05 (0)	0.8864 (0.069)	0.808 (0.156)	0.05 (0)	0.9637 (0.038)
pool_lasso	0.684 (0.14)	0.05 (0)	0.89 (0.069)	0.684 (0.14)	0.05 (0)	0.89 (0.069)	0.804 (0.164)	0.05 (0)	0.9628 (0.046)

	K=4 (n=100)			K=5 (n=150)		
Model	TPR (SE)	FPR (SE)	AUC (SE)	TPR (SE)	FPR (SE)	AUC (SE)
BLINK	0.98 (0.061)	0.05 (0)	0.9972 (0.006)	1 (0)	0.05 (0)	0.9997 (0.001)
mar_ss	0.992 (0.04)	0.05 (0)	0.9982 (0.005)	1 (0)	0.05 (0)	1 (0)
mar_lasso	0.988 (0.048)	0.05 (0)	0.9974 (0.007)	1 (0)	0.05 (0)	0.9997 (0.001)
pool_ss	0.024 (0.077)	0.05 (0)	0.7813 (0.088)	0.848 (0.164)	0.05 (0)	0.9647 (0.047)
pool_lasso	0.044 (0.093)	0.05 (0)	0.7698 (0.089)	0.848 (0.169)	0.05 (0)	0.9648 (0.047)

Table 4.2: Model Comparison Result: True Positive Rate (TPR) and Area Under the Curve (AUC) are shown in this table for all the methods, with the mean calulcated across 50 replicates with False Positive Rate (FPR) being controlled at 0.05. The numbers inside parentheses are the corresponding standard deviation.

4.4.3 Computational Information

As with the computational time, for a single replicate, our linked regression model takes approximately 30 hours with 30000 iterations, using computing server with 96GB memory space and 12 total cores. To implement the multiple replicates, we applied parallel computing using in-house clusters, which took same amount of time. Considering high dimensionality in data sets, complicated prior settings and Bayesian sampling scheme, the computing time is reasonable. Computational time reduction could be implemented by incorporating parallel computing strategy inside the sampling process or applying more advanced programming languages and etc., which will be further investigated in the future.

4.5 Application

We apply our methodology to RPPA based pathway analysis incorporating the genomic and epigenomic upstream factors across multiple cancer types belonging to the same pan-cancer group assuming that they have shared RPPA regulatory mechanisms. Our goal is to apply our methodology framework in detecting the key upstream factors that affect RPPA based pathway activities for multiple cancer types, borrowing strength and learning the similarity mechanisms via analyzing the common factors that are detected by different cancer types.

Our integrative analysis comprises multiple genomic platform profiles as upstream factors including mRNA, microRNA, copy number alteration and methylation. As the outcomes, we focus on pathway activities at protein level, and more specifically, we explored 12 biologically important pathways from the aspect of tumor cell behavior and therapeutic response. As described in TCGA RPPA study by Akbani *et.al* [4], these pathways are: apoptosis, breast reactive, cell cycle, core reactive,

DNA damage response, EMT, PI3K/AKT, RAS/MAPK, RTK, TSC/mTOR, hormone receptor, and hormone signaling (breast). Pathway-Gene membership table is shown in Supplementary materials. Different cancer types can be investigated using our methodology framework, in our application, we focus on multiple cancer groups based on different resources of recent studies. For example, recent study analyzed gynecological and breast cancer types aiming to identify commonality and difference in their features at molecular level through comprehensive analysis of molecular data containing 2579 tumors from TCGA [11]. The study typically focuses on five cancer types: Breast invasive carcinoma (BRCA), Cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC), Ovarian serous cystadenocarcinoma(OV), Uterine Corpus Endometrial Carcinoma (UCEC), Uterine Carcinosarcoma (UCS), which are taken together as "Pan-Gyn" cohort, which we regard it as one of the target cancer groups to be applied with our modeling approach. Similarly, the analysis overview for Pan-Kidney cohort containing Kidney Chromophobe (KICH), Kidney Renal Clear Cell Carcinoma (KIRC) and Cervical Kidney renal papillary cell carcinoma (KIRP), has been published online as in Broad Institute TCGA Genome Data Analysis Center (2016) in public website "http://gdac.broadinstitute.org/runs/analyses_2016_01_28/reports/cancer/KIPAN-TP/index.html". Although each cancer organ site contains specific biological process, we target to study the commonality and distinctions of the RPPA based pathway drivers for each pathway and throughout all three cancer types. Moreover, another pan-cancer group that we focus on is Gastro-Intestinal (GI) cancer group, which defines the group of cancers that influence digestive system. This group includes Stomach adenocarcinoma (STAD), Esophageal carcinoma (ESCA), Colon adenocarcinoma (COAD), Rectum adenocarcinoma (READ). Also, as study shows most recently that integrative analyses of genetic and epigenetic alterations has been implemented for PanCancer Atlas study showing the identifiability in distinguishing

molecular factors of squamous cell carcinomas (SCCs) from other cancer types. The Pan-Squamous includes Head and Neck squamous cell carcinoma (HNSC), Lung squamous cell carcinoma (LUSC), esophageal (ESCA), cervical (CESC), and bladder cancers (BLCA) [81] [82] [83] [83]. Thus, for each cancer group, we explore the key drivers for different pathways, assuming that the cancer types within the same pan-cancer group can have shared common genomic and epigenomic factors that affect the specific RPPA based pathway phenotype.

4.5.1 Data Description

4.5.1.1 RPPA based pathway activity score

To measure the RPPA based pathway activities, tissue specific pathway scores need to be calculated and incorporate into the integrative model as outcomes. Instead of using gene expression data sets, the pathway score is calculated based on RPPA data sets. Considering that the pathway score needs to be unsupervised and yields to single sample in our case. Different methods lie in this category, such as combined z-score, single sample Gene Set Enrichment Analysis (GSEA), Pathway Level analysis of Gene Expression (PLAGE) [8] [62] [109] and gene set variation analysis (GSVA) [35]. We can apply different methods in our study, here, in our framework, we chose to utilize combined z-score based on RPPA data, which is derived based on proteomic expression of the proteins belonging to specific pathway. Thus, for a specific pathway, for the k^{th} cancer, we calculate the RPPA based pathway score as the outcomes denoted as $\boldsymbol{y_k}$ with dimension $n_k \times 1$ where n_k is the number of the samples that have RPPA information.

4.5.1.2 Genomic Upstream Factors

We consider genomic and epigenomic drivers as large molecular covariates that affect the RPPA based pathway activities. The multiplatform genomic data involves mRNA, microRNA, copy number alteration (CNA) and methylation. Consider a specific pathway, we eager to investigate K cancer types, with index k ranging from 1 to K. For the k^{th} cancer, following shows how we retrieve multiplatform genomic data and form it as design matrices for this specific pathway.

- mRNA

  Based on pathway-gene membership table in Appendix B.2, we include the gene expression of the genes that belong to this pathway, we retrieved and applied standardized gene level tumour tissue RNAseq data collected by Illumina and HiSeq that can be downloaded and retrieved through utilizing a recent developed R package "GeneSurvey" that can be accessed online "https://github.com/MD-Anderson-Bioinformatics/GeneSurvey". The mRNA data set can be represented as X_k^{mRNA} with dimension $n_k^{mRNA} \times P_k$ where n_k^{mRNA} denotes the mRNA sample size for the k^{th} cancer and P_k is the number of the genes in this pathway.

- microRNA

  Our microRNA data was retrieved using package "GeneSurvey" as well. To effectively acquire the microRNA values that should be included into the model, we referred to multiMiR R package and database which investigated the microRNA-mRNA interactive associations across disease and drug associations [93]. For each gene, we collected and include the related microRNA as our covariate if the interaction is validated in multiMiR study. The microRNA data matrix can be represented as $X_k^{microRNA}$ with dimension $n_k^{microRNA} \times P_k^{microRNA}$ where $n_k^{microRNA}$ denotes the microRNA sample size for the k^{th} cancer and

$P_k^{microRNA}$ is the number of microRNAs that interact with the genes in the pathway.

- CNA

  For copy number alteration, we retrieved and obtained the gene level standardized summary of CNA (SNP 6 (HG19,no CNV and no sex chromosomes)) for each gene within the pathway, using "GeneSurvey". Normal samples were filtered out, leaving tumour tissue samples, forming our CNA data matrix denoted as X_k^{CNA} with dimension $n_k^{CNA} \times P_k$ where n_k^{CNA} denotes the CNA sample size for the k^{th} cancer and P_k is the number of the genes in this pathway.

- Methylation

  The Methylation data we collected using package "GeneSurvey" was Human Methylation 450 data sets. It was processed from level 2 to level 3. Thus, it is also gene level summarizing dataset. For each gene in the pathway, we retrieved the corresponding CpGs , more specifically, we collected the CpGs of the methylation sites that are located at specific position of the gene body with specific distance from its transcription start site (TSS), eg. annotated as TSS1500 or TSS200. Here in our analysis, we include them all matching with each of the genes. The data matrix can be represented as X_k^{Methy} with dimension $n_k^{Methy} \times P_k^{Methy}$ where n_k^{Methy} denotes the methylation sample size and P_k^{Methy} is the total number of the methylation CpG values.

Thus, in this way, multiplatform data collection is processed. However, given the fact of the heterogeneity in the sample sizes of different genomic profiles and our model requires the consistency of the sample size, we further checked the intersection of the sample sizes and established a new set of datasets with the reduced matched samples. Table B.2 in Appendix B provides the summary of the sample size of different platform for all the TCGA cancer types.

We define the response vector as well as the covariate matrix as $\boldsymbol{y_k}$ and $X_k = \{X_k^{mRNA}, X_k^{microRNA}, X_k^{CNA}, X_k^{Methy}\}$ for the k^{th} stratum, and we implemented our metholodogy using the data sets which have already been processed. The final modeling results were computed based on 30000 MCMC sampling iterations, with 15000 as burn-in samples and another 15000 as the basis for final inference. Signals were detected based on posterior inference, as described in our Simulation Section, we regard the covariates as significant if marginal $PPI > 0.5$.

4.5.2 Modeling Results and Biological Interpretation

The modeling results in our case study contain two main parts. First, we target to find the genomic and epigenomic upstream factors that affect specific RPPA based pathway activities. Second, we focus on measuring how "similar" the RPPA based pathway is for each group of cancer types by investigating the similarity matrix, higher similarity means more common factors shared by these cancer types for a specific pathway. We examine the results group by group and illustrate the results through figures and descriptions with biological interpretation.

To biologically interpret the results, we focus on Pan-Gyn group upstream factor selection and the similarity inference, and we also check the common similar pathways and upstream factors across all cancer groups. The application results for other pan-cancer groups their interpretations could be further investigated biologically similarly.

Regarding with the cancer group specific results, the upstream factor selection for each cancer type is illustrated in Figure 4.6, Figure 4.9, Figure 4.11, and Figure 4.13. The effective similarity among different cancer types within each group is shown in Figure 4.8, Figure 4.10, Figure 4.12 and Figure 4.14. Overall, we noticed

that among all 4 groups, pathway Hormone signaling (Breast) is the one that has the effective similarity in its driving factors across all groups. When checking the results into details, we found that BCL2, which is a gene belonging to this pathway, with its mRNA as the key factor that affect the protein pathway activity across all the cancer types within Pan-Gyn group. GATA3 and INPP4B are another two genes with their gene expression selected to be significant in driving Hormone signaling pathway activity across 3 cancer types BRCA, CESC, and UCEC. The result is in agreement with the literature that GATA3, as a transcriptional activator, is frequently highly expressed by luminal epithelial molecular cells in the breast. Study found that GATA3 is among of the top genes that have low expression in carcinomas and lead to poor clinical outcomes [72]. Also, study showed that GATA3 is a sensitive and specific marker in diagnosing breast and urothelial carcinomas [68], thus, with its importance, it can affect the aforementioned cancer types via influencing their important signaling pathway activities. INPP4B, as discussed in previous studies, was discovered to be tumor suppressor [25] [31] [115], with evidence from various studies in breast, prostate, and ovarian cancers [89] [96] [42]. Particularly, INPP4B was found to suppress epithelial cell transformation in human breast cancer studies [42]. The common Hormone signaling (Breast) pathway with the significant effective similarity among all cancers indicate that similar patterns of the associations between upstream factors mapped with the proteins/genes with Hormone signaling (Breast) pathway can be drawn across all cancer groups, which can further infer that the hidden associations remain similar across all cancer types. The inference could be potentially given further validation and the results will lead to therapeutic target finding across all cancer types that belong to the four groups.

For a specific cancer type, various functional proteomic regulators were selected as clearly illustrated in Figure 4.6 for Pan-Gyn group. For all 12 pathways, different

numbers of regulators are selected for different individual cancer types. In general, there are more microRNAs that are selected due to multiple microRNAs mapping with the same gene. CNA is not frequently selected throughout the pathways, and important CpGs are selected as important, which affect gene expression and indirectly influence proteomics and the signaling pathway activities.

Figure 4.8 illustrates the effective similarity linkage of the upstream factors that affect 12 RPPA pathways. The linkage shown in this figure all have similarity PPI above 0.5, and the width of the linkage corresponds with the similarity strength. Thus, it can be inferred that, for Gyn group containing 4 cancer types, 7 pathways share similar key drivers between at least one pair of the cancers. More specifically, Apoptosis pathway owns effective similarity in shared drivers across all the cancers, indicating that all 4 cancer types have shared drivers that affect this pathway. Further, Figure 4.7 shows that these cancer types share one common gene "BCL2", which reflects biological aspect that BCL2 was the first gene that inhibits apoptosis signaling pathway in any species, and its critical role has been further investigated recently as cancer target in Adams et al., 2018 [1]. We also noticed that receptor tyrosine kinase (RTK) pathway has more effective similarity linkages especially between BRCA and UCEC, with higher magnitude. When checking the cancer and pathway specific upstream factors, it shows that gene expression of EGFR serves as the common factors among BRCA, CESC and UCS cancer types. EGFR serves as a key regulator of RTK signaling pathway and is found to be overexpressed in breast cancer tissues, that leads to accelerated tumor growth [63] [17]. Moreover, the copy number of gene ERBB2 is the common drivers for cancer types BRCA, CESC and UCEC for RTK pathway as well, biologically given that ERBB/HER belongs to one of the subfamilies of RTK, which includes gene ERBB2, with its amplification in the gene copy number, is discovered in leading to overexpression

of cell proliferation in various of cancer types [22] [46] [16]. Specifically, Hormone receptor signaling pathway has large similarity magnitude among all linkages between UCEC and BRCA, indicating larger commonality in their selected regulators. As can be illustrated from the Figure 4.6 and Figure 4.7, they have same two genes ESR1 and PGR, with the expression pointed out to be significant. Study showed that ESR1 amplification is a common mechanism in proliferating breast cancer [44] and in endometrial carcinoma [61]. Most recent research showed that PGR is one of the important transcriptional regulatory key factors associated with breast cancer molecular subtypes for Estrogen Receptor (ER), which serves as an important therapeutic target in breast cancers [92]. Thus, common targets between UCEC and BRCA serve as potential therapeutic targets in activating or deactivating Hormone receptor signaling pathway and its related functional proteomic activities.

4.6 Discussion

In this work, we have established a novel linked regression modeling strategy that addresses the multi-stratum regression problem. This Bayesian hierarchical modeling framework allows flexibility in borrowing strength among multiple strata while doing signal detection for each stratum. This approach has been shown having several key advantages: (1) higher signal detection accuracy compared with other relevant approaches with the benefit of borrowing strength, especially for rare disease having small sample size; (2) capability of estimating the similarity of the common factors that are selected for different strata; (3) signal detection and similarity estimation are processed simultaneously so that it adaptively borrows strength across strata dynamically. We noticed that our method benefits more to stratum with small sample size, which rely more on prior settings, for large sample stratum, the benefit is limited while still within proper range.

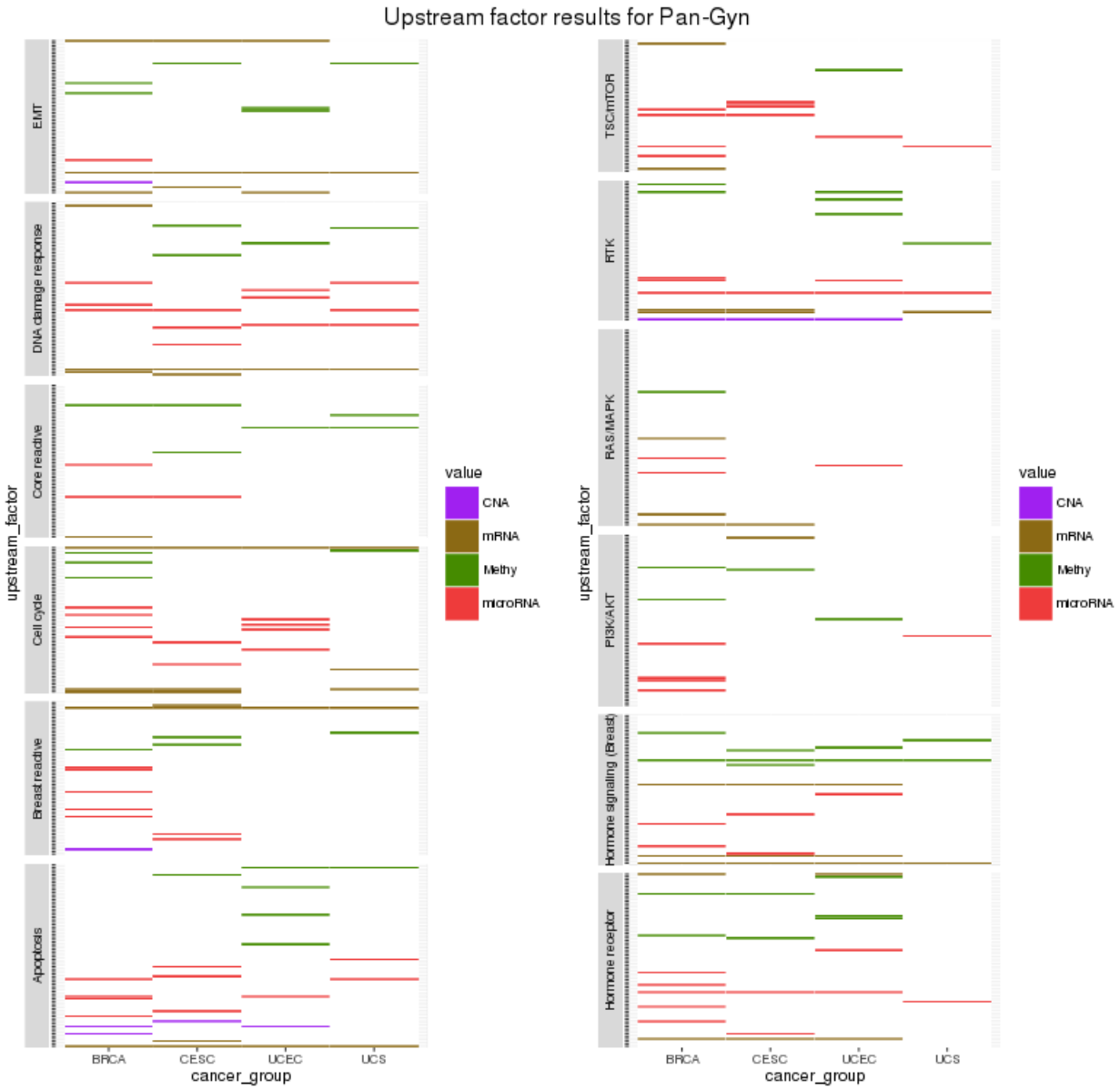


Figure 4.6: Upstream factor detection of Pan-Gyn group: genetic and epigenetic factors that are selection for each cancer and for each pathway using our method are highlighted in different colors as illustrated in the legend.

pathway	platform	BRCA	CESC	UCEC	UCS
EMT	mRNA	CDH11; FN1*; SERPINE1	FN1*; SERPINE1; CLDN7	FN1*; CDH11; SERPINE1	FN1*
	CNA	CLDN7			
	Methylation	cg12805420; cg15698842	cg25020459	cg06875051; cg08051386‡	cg25020459
	microRNA	hsa-miR-142-5p			
DNA damage response	mRNA	CHEK1; CHEK2*; XRCC1	CHEK2*; ATM	CHEK2*	CHEK2*
	CNA				
	Methylation		cg08691422; cg27121267	cg22879270	cg08119584
	microRNA	hsa-miR-233-3p; hsa-miR-20a-5p; hsa-miR-532-5p	hsa-miR-20a-5p; hsa-miR-148a-3p; has-miR-196b-5p	hsa-miR-30a-5p; hsa-miR-361-5p; hsa-miR-1976	hsa-miR-20a-5p; hsa-miR-1976; hsa-miR-532-5p
Core reactive	mRNA	CDH1			
	CNA				
	Methylation	cg15298719	cg15298719; cg11691589	cg08051386	cg26508465; cg08051386
	microRNA	hsa-miR-233-3p; hsa-miR-497-5p	hsa-miR-223-3p		
Cell cycle	mRNA	CCNB1; CCNE1; PCNA*	PCNA*; CCNB1; CCNE1	PCNA*	PCNA*; FOXM1; CCNE1
	CNA				
	Methylation	cg02689825; cg15738421; cg16498681			cg26596863
	microRNA	hsa-miR-423-5p; hsa-miR-20a-5p; hsa-miR-199a-5p; hsa-miR-532-3p	hsa-miR-106b-5p; hsa-miR-194-5p	hsa-miR-15b-5p; hsa-miR-25-3p; hsa-miR-34a-5p hsa-miR-205-5p	
Breast reactive	mRNA	MYH11*	MYH11*; RAB11A	MYH11*	MYH11*
	CNA	GAPDH			
	Methylation	cg06362313	cg11508669; cg03308839		cg08789739
	microRNA	hsa-miR-199a-5p; hsa-miR-93-5p; hsa-miR-185-5p; hsa-miR-324-3p; hsa-miR-629-5p	hsa-miR-107; hsa-miR-130a-3p		
Apoptosis	mRNA	BCL2*	BCL2*; BCL2L11	BCL2*	BCL2*
	CNA	BAD; BCL2L1	BIRC2	BCL2L1	
	Methylation		cg06323957	cg05566965; cg08178356; cg12873919; cg22183731	cg22183731
	microRNA	hsa-miR-125a-5p; hsa-miR-30e-5p; hsa-miR-155-5p; hsa-miR-152-3p	hsa-miR-361-5p; hsa-miR-142-3p; hsa-miR-338-3p	has-miR-155-5p	hsa-miR-30e-5p; hsa-miR-450b-5p

pathway	platform	BRCA	CESC	UCEC	UCS
TSC/mTOR	mRNA	EIF4EBP1; RB1			
	CNA				
	Methylation			cg24863624	
	microRNA	hsa-miR-342-3p; hsa-miR-148b-3p; hsa-miR-324-3p; hsa-miR-181d-5p	hsa-miR-324-3p; hsa-miR-345-5p; hsa-miR-486-5p; hsa-miR-92b-3p	hsa-miR-224-5p	hsa-miR-181d-5p
RTK	mRNA	EGFR; ERBB2	EGFR ; ERBB2		EGFR
	CNA	ERBB2	ERBB2	ERBB2	
	Methylation	cg07474022; cg24384816		cg07474022; cg00915289; cg24657085	cg23757825
	microRNA	hsa-miR-320a; hsa-miR-155-5p*; hsa-miR-224-5p	hsa-miR-155-5p*	hsa-miR-155-5p*; hsa-miR-224-5p	hsa-miR-155-5p*
RAS/MAPK	mRNA	ARAF; JUN	ARAF		
	CNA				
	Methylation	cg22858733			
	microRNA	hsa-miR-199a-5p; hsa-miR-342-5p; hsa-let-7c-5p		hsa-miR-224-5p	
PI3K/AKT	mRNA		PTEN		
	CNA				
	Methylation	cg08874471; cg06466203	cg04707787	cg02313172	
	microRNA	hsa-miR-199a-5p; hsa-miR-425-5p; hsa-miR-140-5p; hsa-miR-181d-5p			hsa-miR-532-5p
Hormone signaling (Breast)	mRNA	BCL2*; GATA3; INPP4B	BCL2*; GATA3; INPP4B	BCL2*; GATA3; INPP4B	BCL2*
	CNA			BCL2L1	
	Methylation	cg00091953; cg25373630*	cg04213746; cg25373630*; cg23756272	cg05356738; cg25373630*	cg15187550; cg25373630*
	microRNA	hsa-miR-146b-3p; hsa-miR-20a-5p	hsa-miR-101-3p; hsa-miR-30a-5p	hsa-miR-532-3p	
Hormone receptor	mRNA	ESR1; PGR	ESR1	ESR1; PGR	
	CNA				
	Methylation	cg20265337; cg25490334	cg13122496; cg25490334	cg27121959; cg03037684; cg03732055	
	microRNA	hsa-miR-15a-5p; hsa-miR-23a-3p; hsa-miR-141-3p; hsa-miR-330-5p; hsa-miR-223-3p	hsa-miR-223-3p; hsa-miR-101-3p	hsa-miR-223-3p; hsa-miR-664a-3p	hsa-miR-181d-5p

Figure 4.7: Table of Upstream factor detection of Pan-Gyn group

We applied our methodology into pan-cancer group analysis, regarding stratum as different cancer types, dependent variable as RPPA based pathway activity scores, and regressors as genetic or epigenetic upstream factors. This particular case study targets to investigate key molecular drivers that influence protein level pathways for each cancer type within a specific group while borrowing strength of the detection across cancer types by assuming that the cancer types in the same pan-cancer category have similarity in their driving upstream factors for a specific pathway. We analyzed 4 pan-cancer groups and 12 important signaling pathways. We identified significant upstream factors for each cancer type and at the same time, we inferred the similarity in common factors for each group for any of the 12 pathways. Thus, our method provides a strategy that investigates common

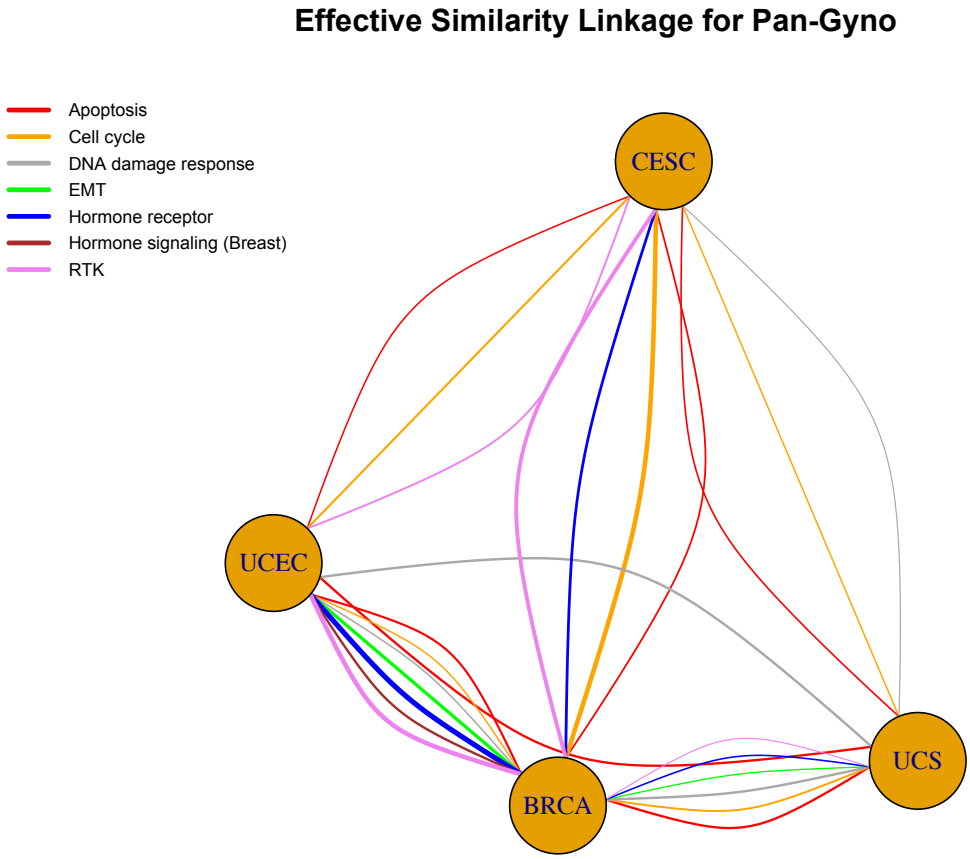


Figure 4.8: The effective similarity linkage for Pan-Gyn throughout the pathways, the pathways listed and illustrated with different colors indicate that 7 pathways select similar upstream regulators. Larger linkage width represents higher similarity.

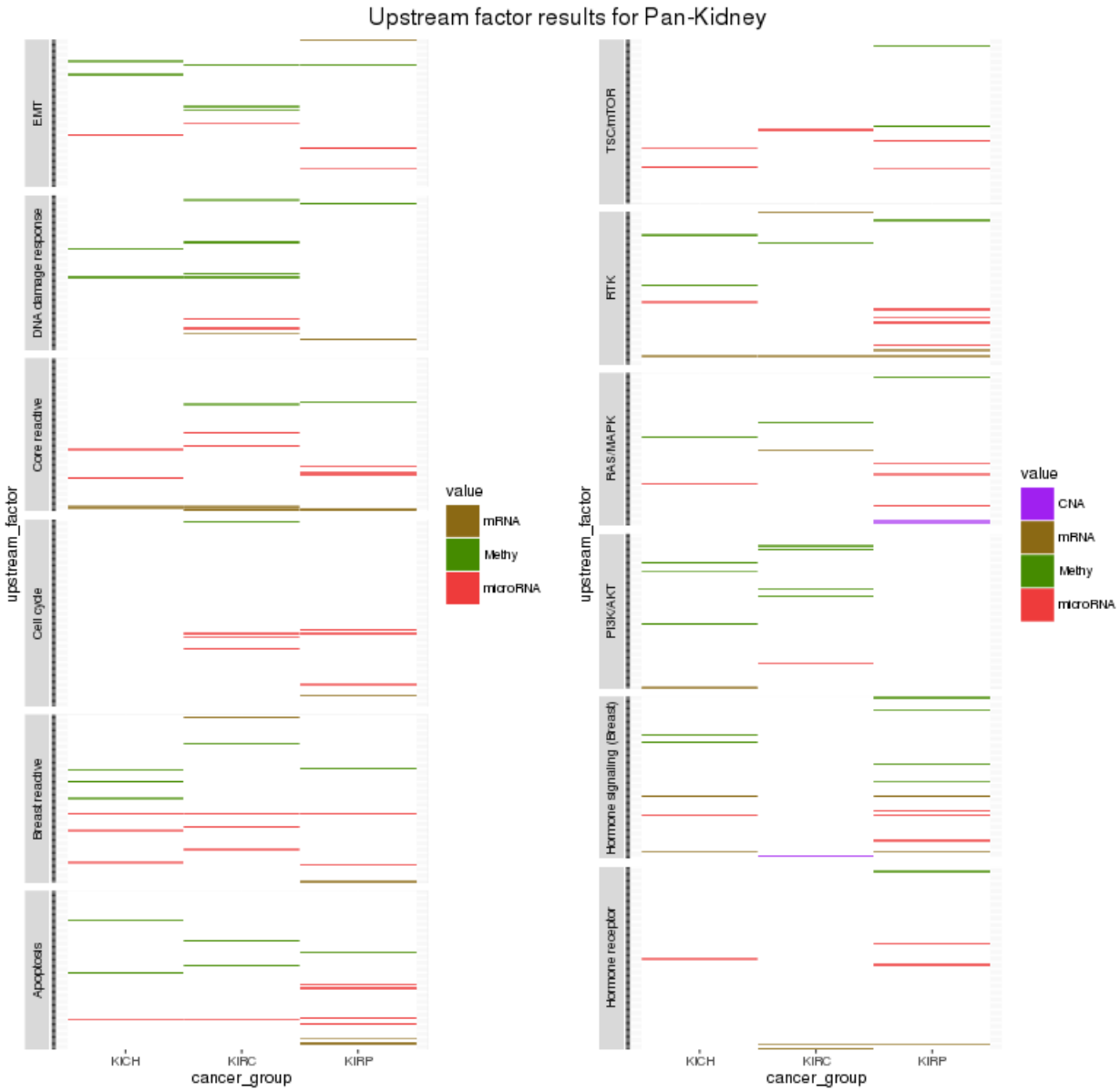


Figure 4.9: Upstream factor detection of Pan-Kidney group: genetic and epigenetic factors that are selection for each cancer and for each pathway using our method are highlighted in different colors as illustrated in the legend.

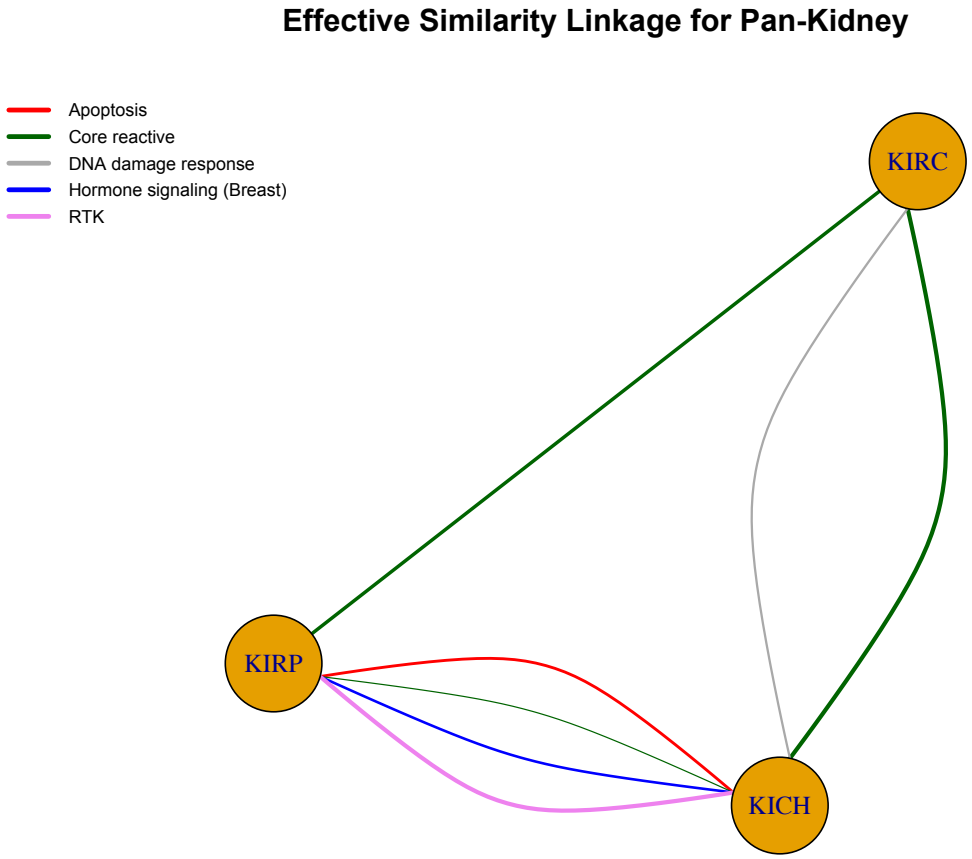


Figure 4.10: The effective similarity linkage for Pan-Kidney throughout the pathways, the pathways listed and illustrated with different colors indicate that 5 pathways select similar upstream regulators. Larger linkage width represents higher similarity.

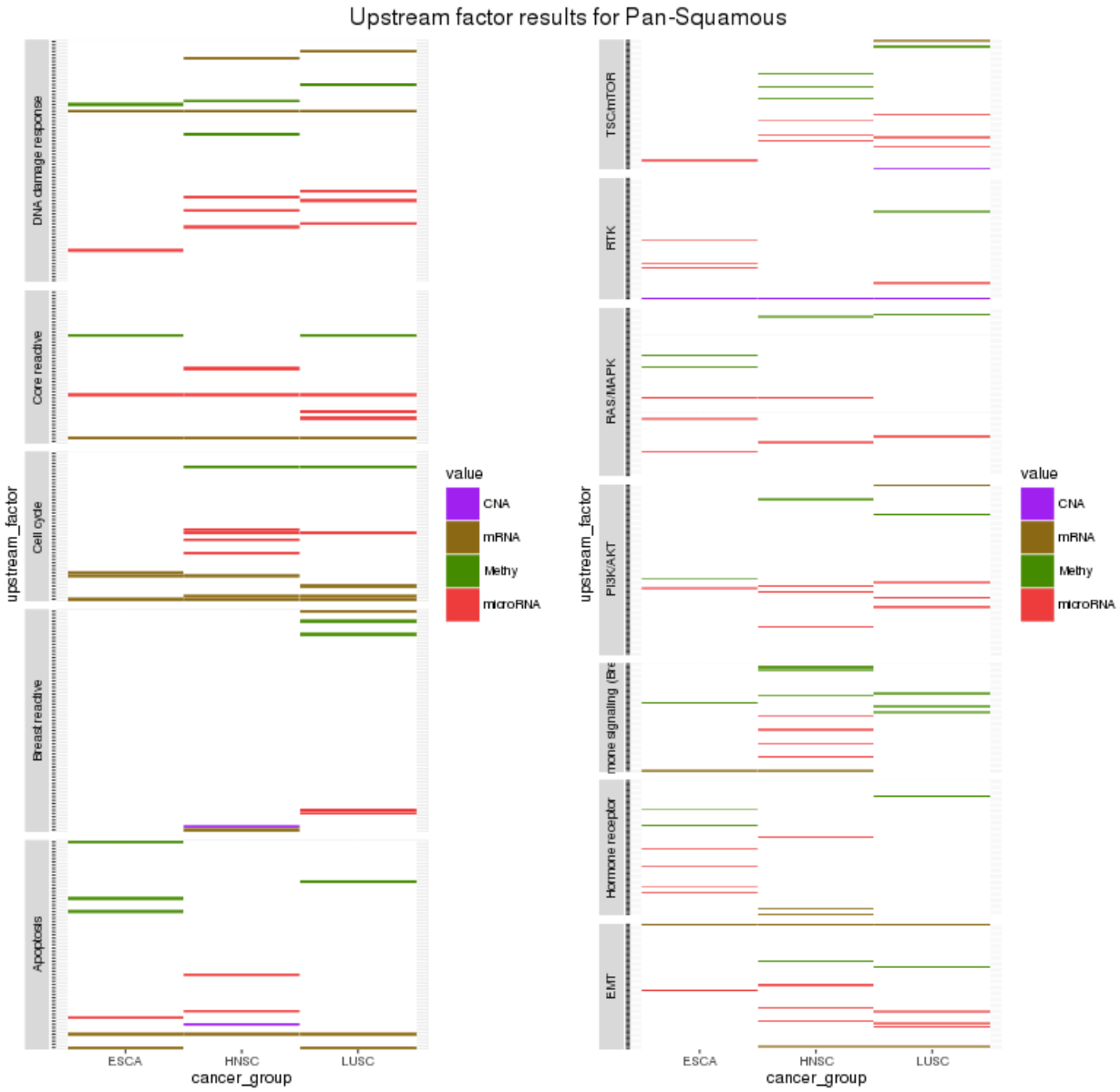


Figure 4.11: Upstream factor detection of Pan-Squamous group: genetic and epigenetic factors that are selection for each cancer and for each pathway using our method are highlighted in different colors as illustrated in the legend.

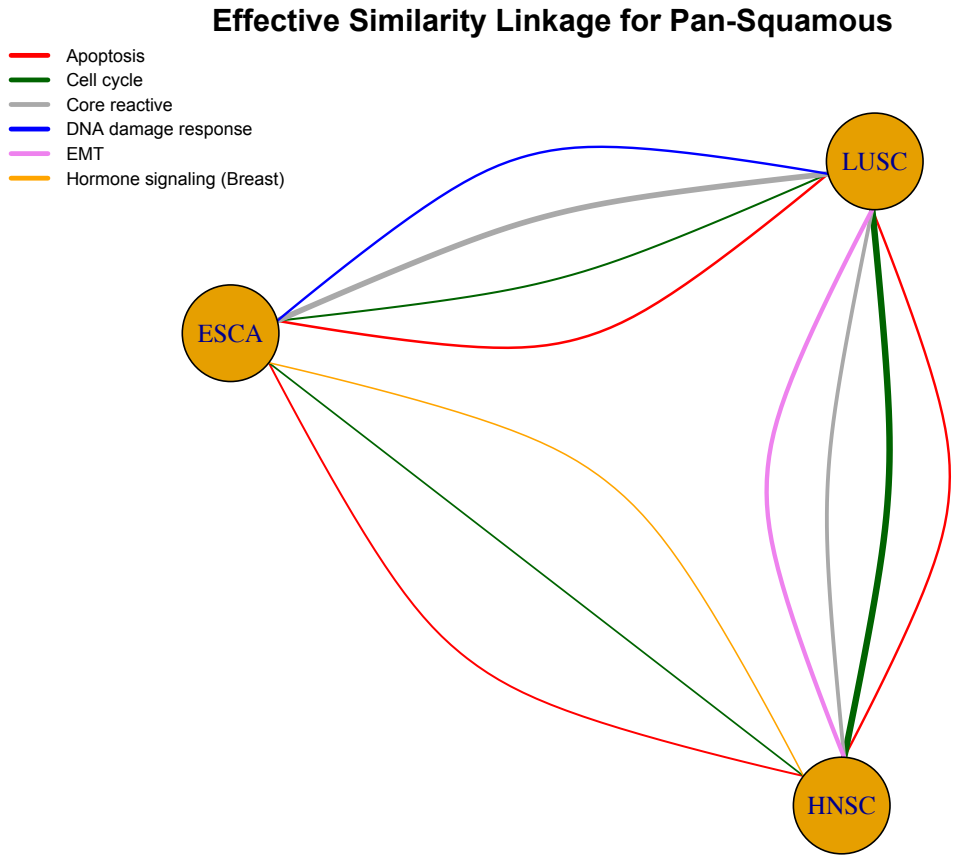


Figure 4.12: The effective similarity linkage for Pan-Squamous throughout the pathways, the pathways listed and illustrated with different colors indicate that 6 pathways select similar upstream regulators. Larger linkage width represents higher similarity.

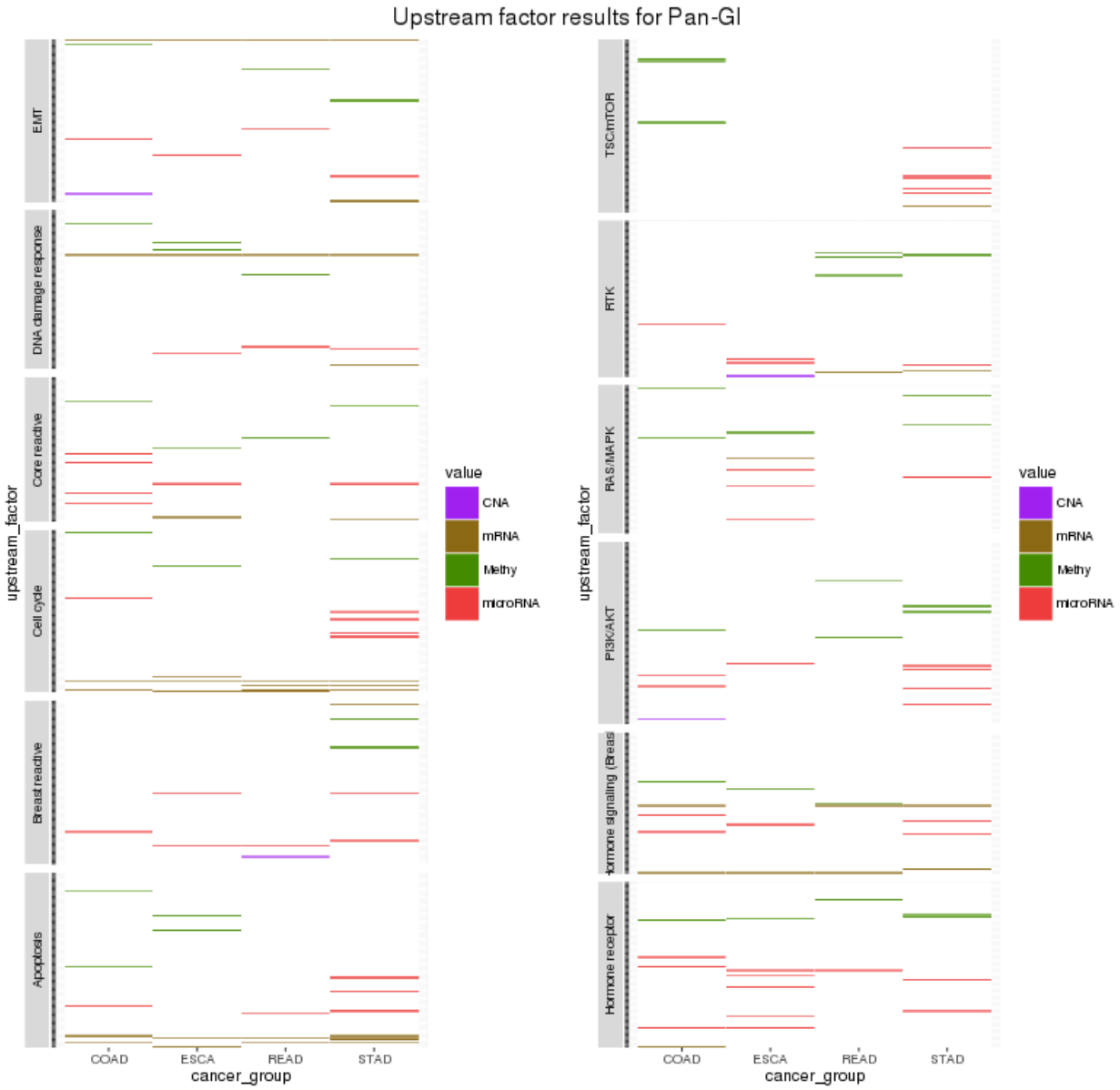


Figure 4.13: Upstream factor detection of Pan-GI group: genetic and epigenetic factors that are selection for each cancer and for each pathway using our method are highlighted in different colors as illustrated in the legend.

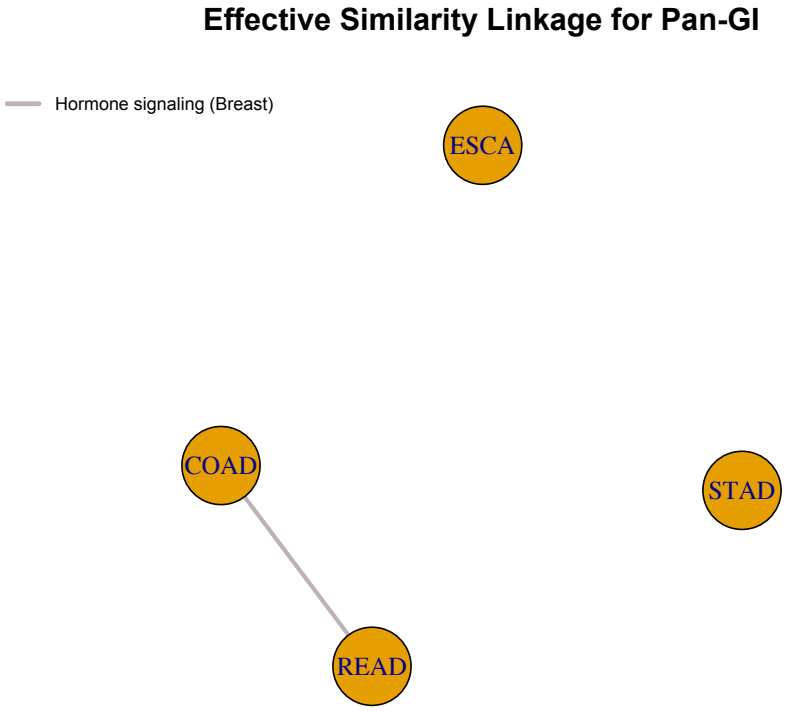


Figure 4.14: The effective similarity linkage for Pan-Gyn throughout the pathways, the pathways listed and illustrated with different colors indicate that only 1 pathways select similar upstream regulators. Larger linkage width represents higher similarity.

molecular drivers that influence some specific pathway activity across cancer types, eg. mRNA BCL2 is the common driver of gynecologic cancer group for Apoptosis pathway, which will potentially serve as an important therapeutic target in clinical field. Similarly, other molecular profiles including CNA, methylation as well as microRNA can also be regarded as pathway specific switches based on our analysis.

In summary, our methodology framework provides a way to link multiple regressions adapted to multiple strata with the same set of covariates, based on the assumption that similarity lies in that some strata share common true significant covariates. Although in our case study, we treat the strata as cancer types, the strata

that we described can take any forms in real practice, such as patients coming from different region, races and age levels. We believe that our methodology benefits the signal detection for each stratum, and similarity inference among multiple strata by capturing if the strata share common factors and to what extent.

However, this project has some weaknesses. First, Given the complexity of the model which simultaneously performs variable selection for each stratum and learns the similarity matrix, the model is not likely to work well in settings for which all strata have very small sample sizes. Another limitation lies in that we do not incorporate interaction term such as considering the association between neighboring genes or CpG sites, or involving modeling nonlinear relationships between pathway activity and genomic upstream factors.

In our framework, possible extensions and generalizations could be implemented in our future work. One extensive direction may be incorporating nonlinearity into our modeling strategy allowing more flexibility in the relationship between dependent variable and covariates. Another direction may be imbedding into group selection for each stratum specific regression allowing horizontal borrowing strength simultaneously. These directions will therefore be of interest in future work.

Chapter 5

Conclusion

In this dissertation, we have introduced novel methods for integrative analysis of omics data with topics ranging from radiogenomics analysis, pathway driver integrative analysis and proteomic based pathway pan-cancer analysis. To start the navigation of the integrative analysis, we started our exploration of radiogenomics which digs into the hidden associations between imaging and genetics incorporating genomic biological mechanisms of multiple platforms using Bayesian hierarchical methodology framework. In Chapter 2, we established a multi-stage framework comprising genomic model which incorporates multiplatform genomic associations, radiogenomic model which investigates the relationships between radiomic features and genomic features. In the process, we dealt with high dimensionality in radiomic features applying Sparse PCA algorithms, which we established Radiomic-meta-Features that can be viewed as radiomic representatives of the original imaging features. In order to detect significant genomic biomarkers associated with RmF and the important RmFs that are related with clinical outcomes, we applied Bayesian adaptive shrinkage method to do features selection, more specifically, we utilized normal gamma prior for the analysis and simultaneously we allowed platform specific effect on each RmF by incorporating platform specific hyperparameters. We applied

Radio-iBAG to a glioblastoma multiforme (GBM) data set from The Cancer Genome Atlas (TCGA), using survival time as the clinical outcome, and identified several potential prognostic genes, genomic platforms as well as radiomic features that implicated in GBM progression.

To explore the integrative analysis of pathway driver detection, we established pathDrive, a new modeling strategy that involves a penalized regression of pathway or gene set scores on multi-platform explanatory variables to identify sparse sets of upstream regulatory factors that can be viewed as key leverage points in the molecular processes underlying the pathway. The pathway scores integrate information across genes within a common network, and provide quantitative summaries that are more functionally relevant than individual gene expressions, and through regression on potential upstream genetic andepigenetic effectors our strategy has the potentiality in identifying key factors in the networks that comprise novel genomic upstream target set. To obtain the sparse molecular upstream target, we applied LASSO with stability selection as feature selection algorithm. We applied our methodology into the analysis of colorectal cancer and incorporated its four specific CMS groups into the selection of the targeted pathways that significantly differentiate CMS subgroups. We finally detect key factors that drive the activities of those pathways in colorectal cancer with biological interpretation.

Motivated by pathDrive analysis, we extended the integrative pathway analysis from single tumor type to multiple tumor types. In Chapter 4, we established a novel analyzing flow of pathway driver detection when there exists cancer types with small sample size, we borrow strength of the signal detection from the ones with enough sample size using Markov Random Field prior setting, which leads to higher power in signal detection. The construction of the prior setting depends on the hidden similar-

ity among cancer types in terms of the selection of the upstream drivers for the same pathway. And simultaneously, the similarity could be measured and estimated from the model. We applied our methodology into different cancer groups: Pan-Kidney, Pan-GI, Pan-Gyn and Pan-Squamous. For each group, we targeted 12 proteomic based pathways and detected their pathway drivers for each cancer type incorporating their selection similarity among different tumor types within that group. Also, the similarity was estimated for each group and given proper biological interpretation.

The proposed framework also leaves substantial space for future research. Following I list some of the limitations and possible directions for improvement:

1. *Radiomics:* In Radio-iBAG project, during the imaging preprocessing, we targeted 2-D pixel based features where they were derived from the imaging slide which had the largest tumor area. However, more sufficient information could be drawn from 3-D volumetric features where they could be calculated from 3-D tumor images combining multiple scanning slices. Another direction in Radiomics part lies in that we can incorporate spatial analysis into radiogenomics framework, directly associating spatial diversity with genomics and clinical outcomes.

2. *Genomics:* We include multiplatform genomic factors in all of the analyzing framework, however, we did not include the relationships and interactions between neighboring or biological related genes and their corresponding upstream factors. Future work can extend the analysis by incorporating this information to delineate a more detailed picture in detecting biological mechanisms, which may involve advanced computing techniques and more complicated biological interpretations.

3. *Pathway:* We investigated into the complicated relationships between path-

way activity and the genetic or epigenetic upstream factors. We focused on detecting linear relationship, however, it could be further extended into nonlinear relationship detection that will provide more sufficient and complex biological processes. From another perspective, when choosing the target pathways, in the case study, the pathways that we aimed were the ones that could significantly differentiate CMS groups. In the future direction, the target pathways could be extracted based on other criterion such as the relatedness with clinical outcomes.

This dissertation mainly focuses on integrative analysis of omics data sets covering multiple areas: radiomics, genomics, proteomics and pathways. It is one of the first efforts in using Bayesian and computational statistical techniques in analyzing multi-platform data resources aiming to get deeper understanding of biological mechanisms and processes. The listed future directions can improve the current study, which will lead to more accurate biomarkers in the context of precision medicine, and better analyzing flow for integrative analysis.

Appendix A

Radio-iBAG Implementation Details

A.1 Full conditional posterior distribution

The general posterior distribution of the coefficient parameter as well as other hyperparameters for the regression model either for the *Radiogenomic Model* or the *Radiogenomic Clinical Model* are shown below.

Consider the linear regression formula: $\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$.

In the *radiogenomic model*, $\boldsymbol{Y}$ denotes the specific RF, X is the matrix of the genomic platform combinations. In the *radiogenomic clinical model*, $\boldsymbol{Y}$ denotes the clinical outcome, X represents the RF combinations modulated by different gene expression parts explained by different genomic platforms. The full posterior distributions are

$$\boldsymbol{\beta}|rest \sim Normal((X^TX + \sigma^2 D_\tau^{-1})^{-1}X^TY, (X^TX + \sigma^2 D_\tau^{-1})^{-1}\sigma^2)$$

$$\sigma^2|result \sim IG(a+n/2, b+(Y-X\boldsymbol{\beta})^T(Y-X\boldsymbol{\beta})/2)$$

$$\psi_{ji}|rest \sim GIG(a=\gamma_j^{-2}, b=\beta_{ji}^2, p=\lambda_j-\frac{1}{2})$$

$$\lambda_j|rest \sim (1/\lambda_j)^{\tilde{a}} exp\{-\tilde{b}\gamma_j^{-2}/(2\lambda_j)-c\lambda_j\} \times \prod_{i=1}^{p_j} \psi_{ji}^{\lambda_j}/\{(\Gamma(\lambda_j))^{p_j}(2\gamma_j^2)^{p_j\lambda_j}\}$$

$$\gamma_j^{-2}|rest \sim Gamma(\tilde{a}+p_j\lambda_j, (\tilde{b}/\lambda_j+\sum_{i=1}^{p_j}\psi_{ji})/2),$$

If applying to the *radiogenomic clinical model*, j denotes the RF combination that are modulated by the gene expression that is explained by the j^{th} platform, and k represents the k^{th} RF; if applying to the *radiogenomic model*, j is the genomic platform type index, i is the gene index.

Specifically, λ_j is sampled through the Metropolis-Hastings method, the proposed family is $exp(\sigma_\lambda^2 z)\lambda_j$, and z comes from the standard normal distribution. The acceptance rate is controlled between 20% and 30%.

A.2 Data Preprocessing for GBM

In this section, we mainly present data preprocessing techniques and procedures, especially for GBM MRI imaging data, including imaging preprocessing, feature extraction and description. Further, we illustrate our preliminary checking results for radiomic features (RFs) and describe how radiomic-meta-features (RmFs) are generated. For the genomic platform data sets, we perform missing value imputation; detailed information is described in the last subsection.

A.2.1 Radiomic-feature preprocessing

Our GBM MRI imaging data set was downloaded from The Cancer Imaging Archive with two imaging modalities: T1-post contrast and T2-weighted FLAIR.

In brief, the dicom MRI images were converted to the nifti format (.nii) using MRIConvert software ($http://lcni.uoregon.edu/ \sim jolinda/MRIConvert/$). We then performed pre-processing steps on the MRI images by following a certain pipeline, and further obtained the 3D tumor volumes. The pipeline is described here.

a. Non-uniformity correction: We used a nonparametric intensity non-uniformity normalization (N3) correction module in MIPAV (v 6.0) [70] to correct the shading artifacts that resulted from partial volume averaging errors of the MRI instrument. The N3 algorithm iteratively estimates the true tissue intensity distribution, as the shading artifacts lead to reduced signal intensities in certain image regions.

b. Registration: We used medical image processing, analysis and visualization (MIPAV) software to register the T2-FLAIR N3 corrected images to the respective T1-POST N3 corrected images. We used the normalized mutual information criterion in the optimized automatic registration module in MIPAV ((http://mipav.cit.nih.gov/).

c. Segmentation: Semi-automated segmentation of the tumor regions in 3D was performed by our clinical experts using the Medical Image Interaction Toolkit 3M3 (http://www.mint-medical.de/). This software features slice-by-slice contour drawing, correction tools and 3D interpolation of the tumor region, which are utilized to perform the tumor segmentation.

d. Re-slicing: Lastly, we re-sliced the original as well as the pre-processed images to widths of 1 millimeter using the NIFIT toolbox in MATLAB ($http://research.baycrest.org/ \sim jimmy/NIfTI$). Image Feature Computation: Textural and regional image features. Computer-based texture analysis of medical images

such as MRI scans depicts some quantitative properties like the measurement of regional intensity variations. Using these properties in the textural analysis can further help in the prediction of several factors such as molecular subtypes of the disease and patient survival times.

In this study, the image-derived textural features are utilized in analyzing the relationship between imaging data, genomic multiplatform data sets, and clinical outcomes. For this analysis, we derived textural features from the largest axial 2D slice of the tumor area [122]. These textural features were obtained from the following two-step process.

1. Image filters: We used Laplacian of Gaussian (LoG) [36] and Gaussian filters [37] to filter the MR images at five different scales, so as to obtain fine, medium and coarse transforms of the 2D tumor region. The LoG filter, commonly used for edge detection, is a measure of the second spatial derivative of an image. The Laplacian filter is applied to an image that has been smoothed using a Gaussian filter; hence, reducing the sensitivity to noise. We used five standard deviations (σ) to derive fine features at 0.2mm and 0.4mm, medium features at 1.5mm and 2.5mm, and coarse features at 5mm.

2. Texture features and summary measures: In terms of textural features, we calculated Haralick features for both T2-weighted FLAIR and T1-weighted post contrast images. The gray-level-co-occurrence matrices (GLCMs) were derived for both the original and pre-processed images. Further, from the GLCMs, we computed Haralick statistical features at 4 different distances (1mm, 2mm, 4mm and 8mm). Besides the textural features, we obtained some summary features such as the mean intensity of the images, entropy and uniformity measures. These features are also

known as "TxR" features [28]. In addition to the textural features, we computed the area and mean intensity of the largest tumor slice. These are referred to as the "regional" features.

After the extraction of the features, we conducted further normalization by calculating two different ratios: type-1 and type-2. "Ratio type-1" corresponds to the ratio between the features computed at different filters and the features computed from the original images (unfiltered features). "Ratio type-2" corresponds to the ratio between features computed at the coarsest scale and those computed at the finer scale.

In our analysis, we considered three major types of RFs: Haralick features, histogram features and regional features. We have 13 total Haralick features, with the names and corresponding calculation formulas listed in Table A.1. Histogram features are calculated from the histogram of the image intensity distribution, with a total of 3 typical features, including the mean intensity, entropy and uniformity characteristics, with formula shown in Table A.2. As mentioned in the imaging preprocessing section, the two regional features that we utilized were the area and mean intensity of the largest tumor slice.

A.2.2 RF and RmF description

In total, we have 972 RFs that we extracted from the original imaging data. Considering that some features were extracted from the same imaging modality, or were processed using the same algorithm, we categorized 972 RFs into 20 groups based on their properties, extracting procedure and imaging modalities. Table A.3 illustrates the name of the RF groups and the corresponding brief description.

Table A.1: Haralick features and formulas

Textural Features	**Formula**
Energy	$f_1 = \sum_i \sum_j \{p(i,j)\}^2$
Contrast	$f_2 = \sum_{n=0}^{N_g-1} n^2 \{ \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j) \vert\vert i-j\vert = n\}$
Correlation	$f_3 = \frac{\sum_i \sum_j (ij)p(i,j) - \mu_x \mu_y}{\sigma_x \sigma_y}$
Sum of squares: variance	$f_4 = \sum_i \sum_j (i-\mu)^2 p(i,j)$
Inverse difference moment (local homogeneity)	$f_5 = \sum_i \sum_j \frac{1}{1+(i-j)^2} p(i,j)$
Sum average	$f_6 = \sum_{i=2}^{2N_g} i p_{x+y}(i)$
Sum variance	$f_7 = \sum_{i=2}^{2N_g} (i - f_8)^2 p_{x+y}(i)$
Sum entropy	$f_8 = -\sum_{i=2}^{2N_g} p_{x+y}(i) log\{p_{x+y}(i)\}$
Entropy	$f_9 = -\sum_i \sum_j p(i,j) log\{p(i,j)\}$
Difference variance	f_{10} =variance of p_{x-y}
Difference entropy	$f_{11} = -\sum_{i=0}^{N_g-1} p_{x-y}(i) log\{p_{x-y}(i)\}$
Cluster shade	$f_{12} = \sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} \{i+j-\mu_x-\mu_y\}^3 p(i,j)$
Cluster prominence	$f_{13} = \sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} \{i+j-\mu_x-\mu_y\}^4 p(i,j)$

* N_g denotes the number of distinct grey levels used; μ_x,μ_y,σ_x and σ_y are the means and standard deviations of the partial probability density p_x and p_y; $p_x(i) = \sum_{j=1}^{N_g} p(i,j)$, $p_y(j) = \sum_{i=1}^{N_g} p(i,j)$; $p_{x+y}(k) = \sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} p(i,j), i+j=k$ and $k = 2,3,...,2N_g$; $p_{x-y}(k) = \sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} p(i,j), |i-j| = k$ and $k = 0,1,...,N_g-1$.

Table A.2: Histogram Features and formulas

Histogram Features	**Formula**
Mean	$m = \sum_{i=0}^{L-1} z_i p(z_i)$
Uniformity	$U = \sum_{i=0}^{L-1} p^2(z_i)$
Entropy	$e = -\sum_{i=0}^{L-1} p(z_i) log_2^{p(z_i)}$

* z_i denotes a random variable indicating intensity, $p(z_i)$ is the histogram of the the intensity values within ragion of interest (ROI), and L is the number of possible intensity levels.

The 972 preprocessed RFs are highly correlated when checking for Pearson correlation. Figure A.1 clearly shows that the RFs are highly correlated with the block structure due to two imaging modalities as well as two major normalizing approaches (ratio-1 and ratio-2) applied during feature extraction.

In our analysis, we chose to apply dimensional reduction approaches to our RFs. The

Table A.3: RF groups and description

RF group name	Description
F_Region	Regional features including tumor area, maximum intensity, minimum intensity, mean intensity of T2-weighted FLAIR image
F_LoG_Tex_R1	Ratio1(filter/unfiltered) Haralick texture features derived from LoG filtered T2-weighted FLAIR image
F_Unft_Hist	Histogram features derived from unfiltered T2-weighted FLAIR image
F_LoG_Hist_R1	Ratio1(filter/unfiltered) histogram features derived from LoG filtered T2-weighted FLAIR image
T1_Region	region features including tumor area, maximum intensity, minimum intensity, mean intensity of T1-post contrast image
T1_LoG_Tex_R1	Ratio1(filter/unfiltered) Haralick texture features derived from LoG filtered T1-post contrast image
T1_Unft_Hist	Histogram features of unfiltered T1-post contrast image
T1_LoG_Hist_R1	Ratio1(filter/unfiltered) histogram features of LoG filtered T1-post contrast image
F_LoG_Tex_Fine	Haralick features derived from LoG filtered FLAIR image with fine scale
F_LoG_Tex_R2	Ratio2(coarse/fine) haralick features derived from LoG filtered T2-weighted FLAIR image
F_LoG_Hist_Fine	Histogram features of fine LoG filtered T2-weighted FLAIR image
F_LoG_Hist_R2	Ratio2(coarse/fine) histogram features of LoG filtered T2-weighted FLAIR image
F_Gauss_Hist_Fine	Histogram features of Gaussian filtered FLAIR image with fine scale
F_Gauss_Hist_R2	Ratio2(coarse/fine) histogram features derived from Gaussian filtered T2-weighted FLAIR image
T1_LoG_Tex_Fine	Haralick features derived from LoG filtered T1-post contrast image with fine scale
T1_LoG_Tex_R2	Ratio2(coarse/fine) haralick texture features derived from Gaussian filtered T2-weighted FLAIR image
T1_LoG_Hist_Fine	Histogram features of LoG filtered T1-post contrast image with fine scale
T1_LoG_Hist_R2	Ratio2(coarse/fine) histogram features derived from LoG filtered T1-post contrast image
T1_Gauss_Hist_Fine	Histogram features of Gaussian filtered T1-post contrast image with fine scale
T1_Gauss_Hist_R2	Ratio2(coarse/fine) histogram features derived from Gaussian filtered T1-post contrast image

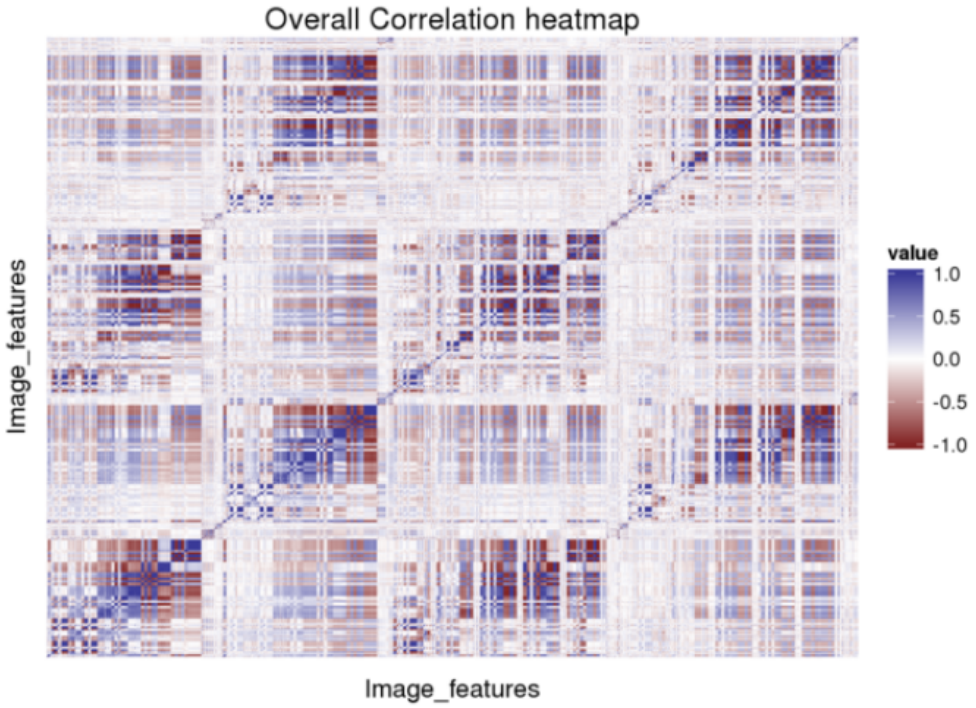


Figure A.1: Heatmap of the Pearson correlation among radiomic features (972 features)

typical technique is principal component analysis (PCA), however, the key limitation lies in that it does not lead to sparse loadings, making it harder to interpret the results. Hence, alternatively, the sparse PCA algorithm developed by [125] was applied with the formulation described below.

Suppose Z_i is the i^{th} principal component derived from ordinary PCA of matrix $\boldsymbol{X}$ with n samples and p predictors, where the loading matrix is denoted as V_i. By regressing PC on X with penalization, sparse loading can be achieved.

$$\hat{\beta} = \arg_{\beta} min||Z_i - \boldsymbol{X}\beta||^2 + \lambda||\beta||^2 + \lambda_1||\beta||_1$$

where $||\beta||_1 = \sum_{j=1}^{p} |\beta_j|$. The updated sparse loading can be expressed as $\hat{V}_i = \frac{\hat{\beta}}{||\hat{\beta}||}$, and $\boldsymbol{X}\hat{V}_i$ is the i^{th} approximated principal component. For a more detailed theorem and proof, see the appendix for the publication by Zou and Hastie (2006).

The loadings as well as the leading principal components were derived from both ordinary PCA and sparse PCA, based on RF-prespecified groups, the squared loading proportion of the principal analysis is calculated respectively. This information is shown in Figure A.2 below and in Figure 2.2 in the main text. When comparing these two heatmaps, we can explicitly see a great difference in the sparsity level. We utilized 22 leading principal components derived from sparse PCA as our imaging features in modeling stage II and stage III. We call them "radiomic-meta features" (RmFs) in the analysis.

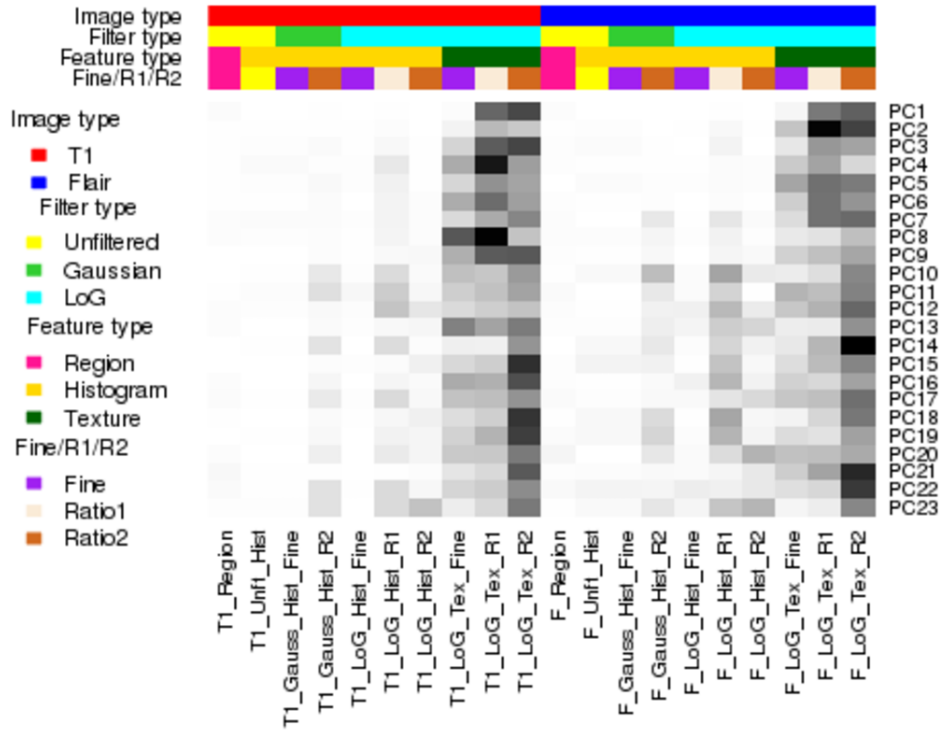


Figure A.2: PCA Squared loadings proportion for each RF group. For each of the 23 PC scores, the sum of the squared loadings of each group is calculated after dividing by the total sum of the squared loadings that equals exactly 1. The heatmap shows this values in grey level, interpreted the RF group importance for each PC component. The grey level ranging from white to black matches the proportional values ranging from 0 to 1.

A.2.3 Dataset Sample Size

Figure A.3 shows the diagram of the sample size description of genomic data of different platform, radiomic data sample size and the clinical data sample size as well as their intersection sample sizes.

mRNA: continuous data

CN: log transformed continuous data

microRNA: continuous data

Radiomic data: continuous

Clinical: continuous data (we took the survival in month with log2 transformation as the outcomes)

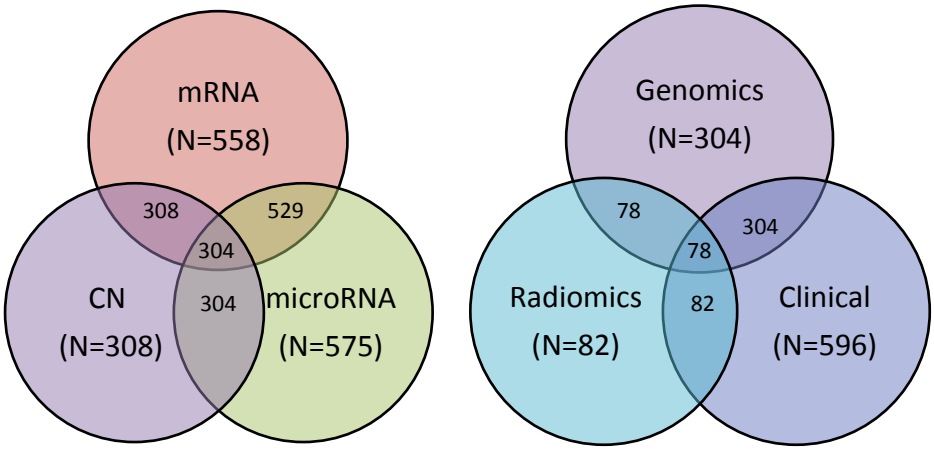


Figure A.3: Diagram of sample size, there are 304 samples having all mRNA, CN and microRNA information, and 78 samples have clinical and Radiomic information.

A.2.4 Missing value imputation for genomic platform data

We have missing values in the copy number data set, with 4.3% of the data missing. To impute the missing values, we chose to use the following steps. First, we impute each NA with the average values of the other patients (mean imputation). Second, using the complete matrix with mean imputation to calculate the correlation matrix between markers, for each target marker with missing elements, we select 3 markers that are the most highly positively correlated with this marker. Third, we regress the target marker on the 3 selected markers and obtain the predicted values. Lastly, we replace the predicted values for the missing elements of the target marker in the original matrix.

A.3 Nonlinearity Checking for Genomic Model

We applied generalized additive model (GAM) in *Genomic Model* for each gene given that GAM, compared with General Linear Model (GLM), can achieve higher flexibility in modeling the genomic mechanisms. To check the existence of nonlinearity, we show the comparison of GLM and GAM in terms of ANOVA p-values for model comparison across all 49 genes. Moreover, we also show the fitted smooth curves and the corresponding confidence interval lines for several genes and the platforms in below.

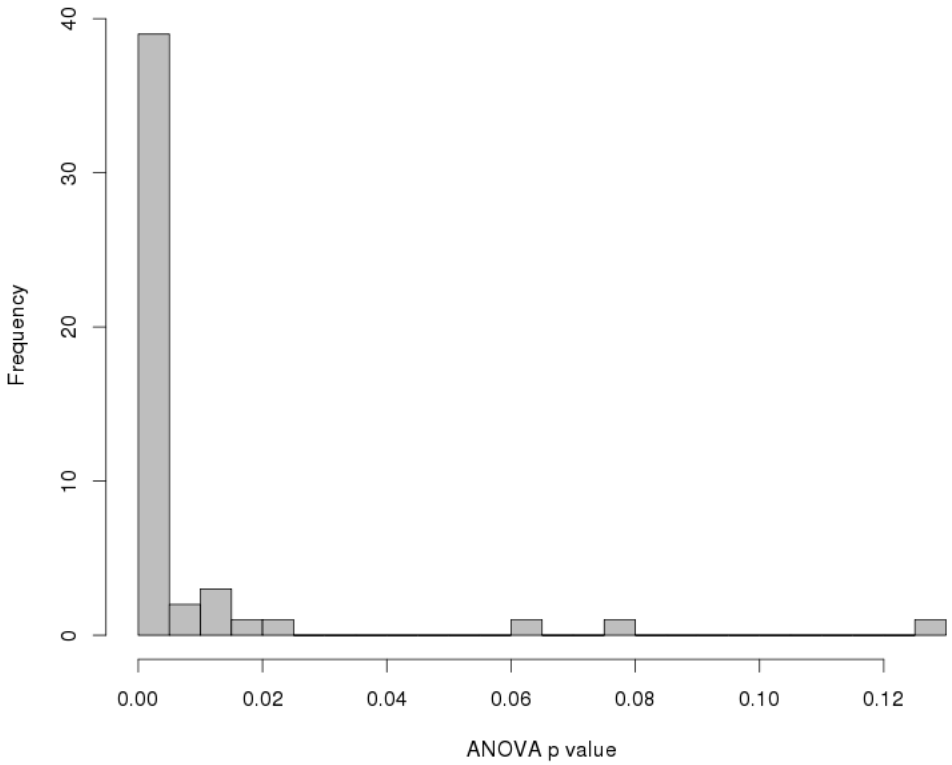


Figure A.4: Histogram of ANOVA p-value when doing model comparison (GLM vs. GAM and GLM is nested into GAM), small p-value indicates the two models have significant difference and GAM is preferred over GLM.

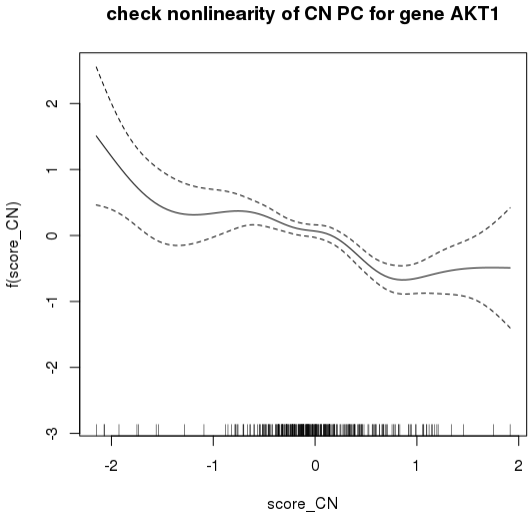


Figure A.5: Fitted smoothing curve for the 1st leading PC score of copy number alteration of gene AKT1, the figure shows the existence of nonlinearity

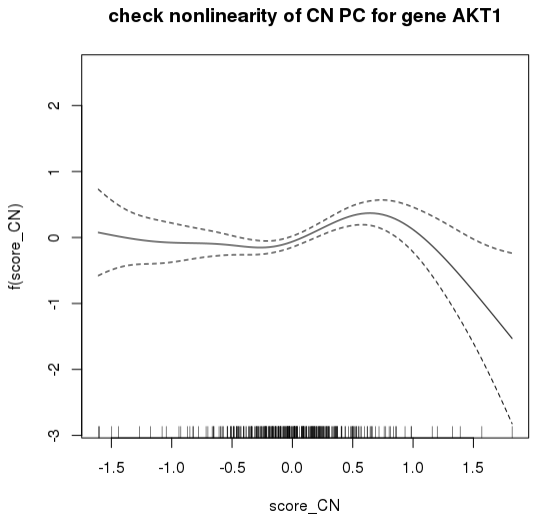


Figure A.6: Fitted smoothing curve for the 2st leading PC score of copy number alteration of gene AKT1, the figure shows the existence of nonlinearity

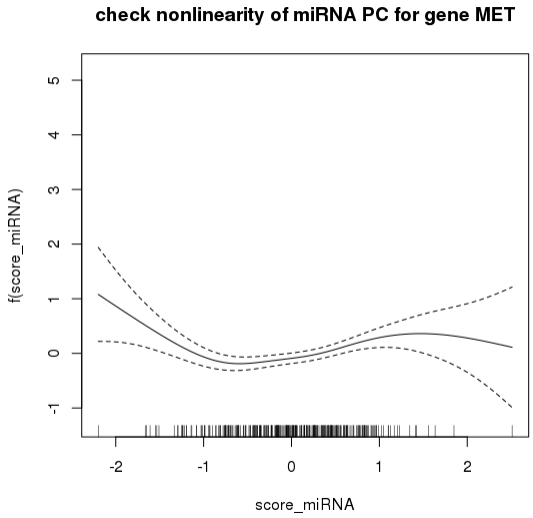


Figure A.7: Fitted smoothing curve for the 1st leading PC score of microRNA mapped with gene MET, the figure shows the existence of nonlinearity

A.4 Additional Results

A.4.1 Magnitude in Stage II

In the *radiogenomic model*, we dig into the relationship between the multiplatform genomic data and RmFs; the magnitudes (posterior mean of the coefficients) are shown in Figure A.8.

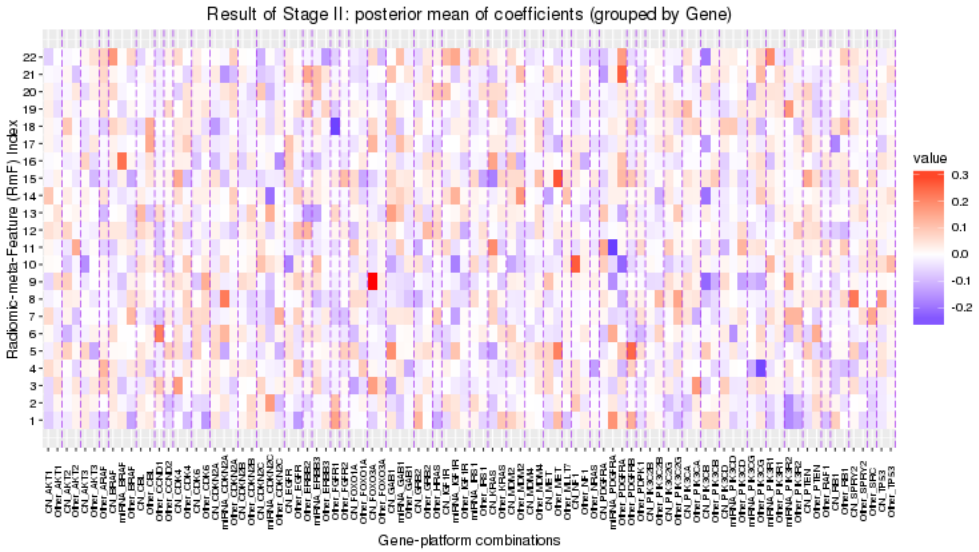


Figure A.8: Posterior mean of magnitude from stage II (radiogenomic model). For each RmF, the posterior mean of β_{jg} is the magnitude of the g^{th}'s mRNA part explained by the j^{th} genomic platform. After filtering, 92 gene-platform combinations are sorted and grouped by gene, and the positive and negative effects are respectively illustrated in red and purple.

A.4.2 Convergence Checking

We applied Bayesian Normal Gamma shrinkage model taking account of the multi-scale datasets in both stage II and stage III, thus, we check the convergence here for the parameters and hyperparameters in stage II and stage III.

In stage III, we have totally 185 parameters including β_{ji}, ψ_{ji} and $j = 1, 2, 3, 4; i =$

$1, 2, 3, ..., 22$, σ^2, λ_j $(j = 1, 2, 3, 4)$ and $1/\gamma_j^2$ $(j = 1, 2, 3, 4)$. We ran MCMC for 30000 iterations and summarize the results using 20000 samples with the burn-in samples removed. We evaluate the convergence by checking traceplots of 3 main parameter vectors: βs, ψs and σ^2, as well as the ratio of $\lambda_j/(1/\gamma_j^2)$ which identifiably leads to the estimation of ψs.

In addition, we check the convergence of the parameters in Stage III using Geweke test [30] where we test for the mean difference between first 10% proportion and the last 50% of the samples for all the parameters. P values are illustrated via histogram plots as in the following figures. The histogram is not skewed to the right which indicates proper mixing for MCMC iterations. In all, our result summary is based on the chains which are long enough to guarantee the convergence.

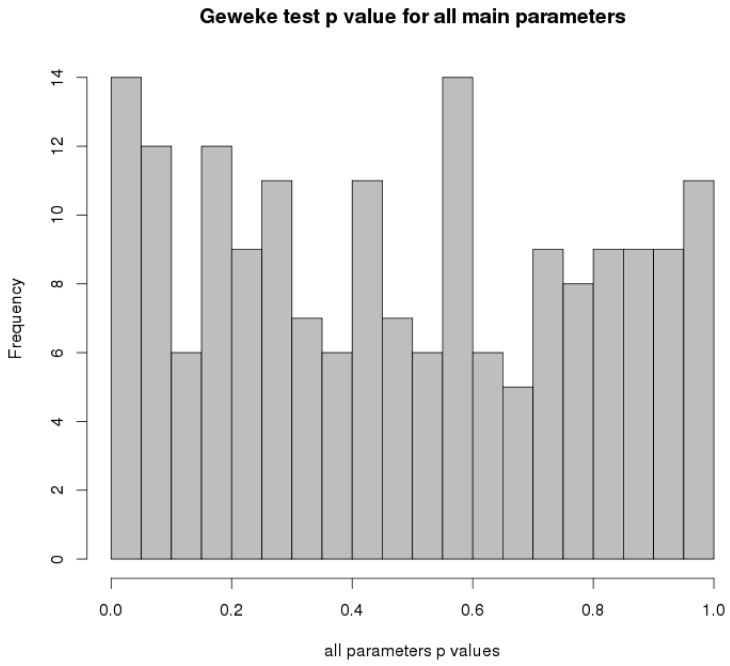


Figure A.9: Geweke test for all parameters, 7.7% have p value smaller than 0.05.

A.5 Sensitivity Analysis

Based on the prior settings described in the paper, for stage III, $\boldsymbol{\mathcal{I}} = \{\boldsymbol{\mathcal{I}_{CN}}, \boldsymbol{\mathcal{I}_{miR}}, \boldsymbol{\mathcal{I}_{O}}, \boldsymbol{\mathcal{I}_{\bar{g}}}\}$, and the effect parameter $\boldsymbol{\alpha} = \{\boldsymbol{\alpha_1}, \boldsymbol{\alpha_2}, \boldsymbol{\alpha_3}, \boldsymbol{\alpha_4}\}$, then the model and prior construction can be expressed as

$$\boldsymbol{Y} = \boldsymbol{\mathcal{I}\alpha} + \boldsymbol{\epsilon}$$

$$\boldsymbol{Y} \sim Normal(\boldsymbol{\mathcal{I}\alpha}, \sigma^2 \boldsymbol{I_{N_{\mathcal{GIC}}}})$$

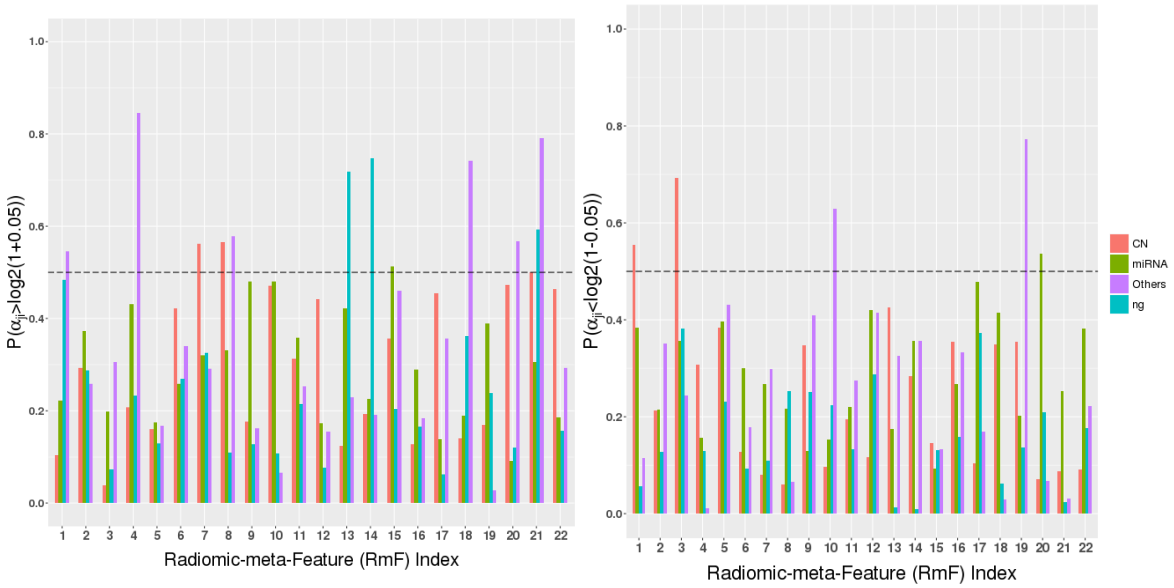
$$\boldsymbol{\alpha} \sim Normal(\mathbf{0}, D_\psi)$$

$$D_\psi = diag(\psi_{1,1}, \psi_{1,2}, ..., \psi_{1,K}, \psi_{2,1}, \psi_{2,2}, ..., \psi_{2,K}, ..., \psi_{J,1}, \psi_{J,2}, ..., \psi_{J,K}),$$

where J denotes the total number of different RF combination types ($j = 1, 2, 3, ..., J$, our $J = 4$), k denotes the RF index ($k = 1, 2, 3, ..., K$). Further, we assign our prior and hyper-prior distributions as $\psi_{j,k} \sim Gamma(\lambda_j, 1/(2\gamma_j^2))$, $\sigma^2 \sim InverseGamma(u_1, u_2)$, $\lambda_j \sim exp(d)$, and $1/(2\gamma_j^2) \sim Gamma(\tilde{e}, \tilde{f}/(2\lambda_j))$.

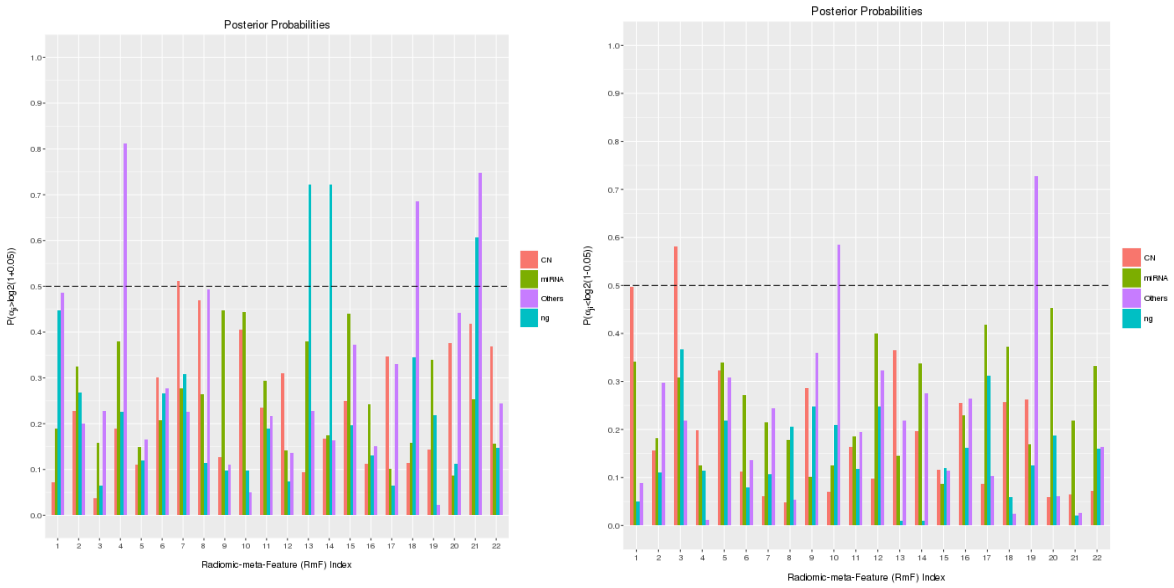
We have hyperparameters d, $\tilde{e}$ and $\tilde{f}$. For hyperparameter $\tilde{f}$, it is suggested that $\tilde{f}$ comes from minimum-length least squares (MLLS) of the coefficients. Thus, we do the sensitivity analysis by adjusting d, $\tilde{e}$. In our analysis, we set up $d = 1$ and $\tilde{e} = 2$, in the sensitivity analysis, we set up $d = 0.5, 2$ and $\tilde{e} = 1, 4$ respectively.

Results in below show that the selected RmF components are almost the same across different hyperparameter settings. Even for some of the features which are not selected given lower value of d, they are close to the margin. Thus, our model is consistent with the hyperparameter settings. Since we applied the same prior for stage II, similar results will be drawn.



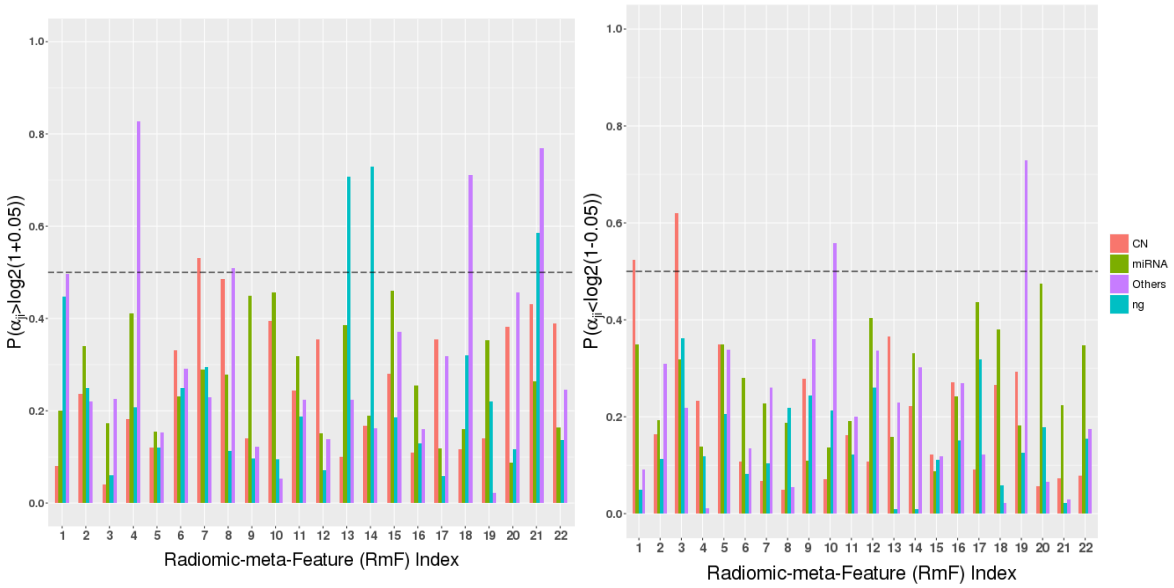
(a) positively significant RmF combinations **(b)** negatively significant RmF combinations

Figure A.10: Results of Stage III when $d = 0.5$



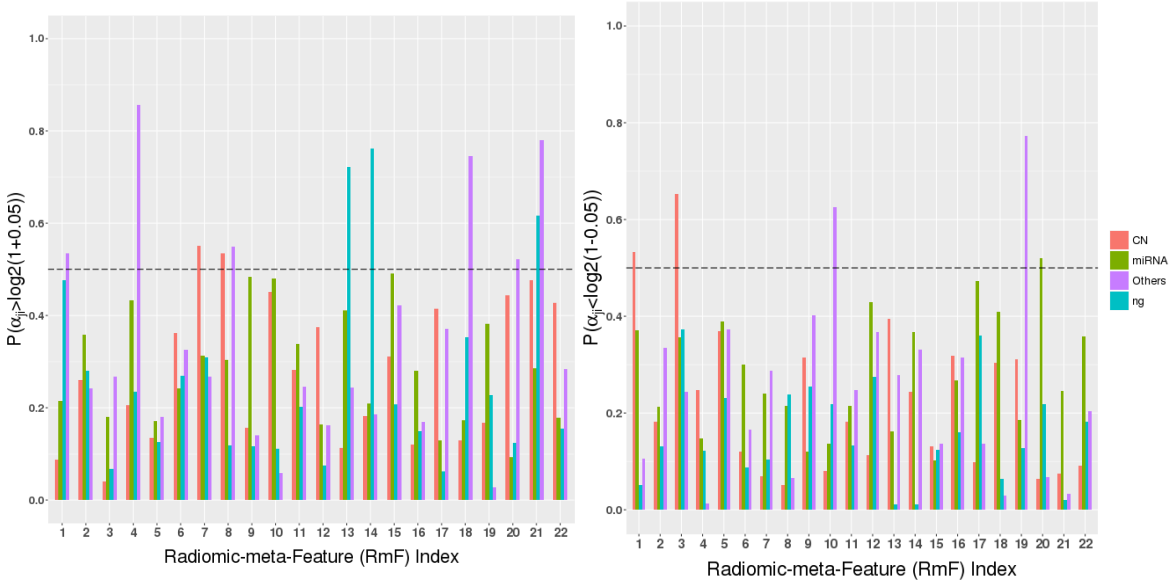
(a) positively significant RmF combinations **(b)** negatively significant RmF combinations

Figure A.11: Results of Stage III when $d = 2$



(a) positively significant RmF combinations **(b)** negatively significant RmF combinations

Figure A.12: Results of Stage III when $\tilde{e} = 1$



(a) positively significant RmF combinations **(b)** negatively significant RmF combinations

Figure A.13: Results of Stage III when $\tilde{e} = 4$

Appendix B

BLINK Implementation Details

B.1 MCMC SAMPLING

B.1.1 Updating $\boldsymbol{\beta_k}$ and $\boldsymbol{\gamma_k}$

In our analysis, we have K regressions, for simplicity, we take one as the example, considering the k^{th} regression:

$$\boldsymbol{Y_k} = X_k\boldsymbol{\beta_k} + \boldsymbol{\epsilon_k}, \boldsymbol{\epsilon_k} \sim \mathcal{N}(\mathbf{0}, \sigma_k^2\boldsymbol{I}) \tag{B.1}$$

Note that the response vector has been centered and the predictor matrix is standardized, so as to avoid intercept and to obtain more interpretable results especially for coefficients. Given the prior described in equation (2) and (3), with the updated prior inclusion probability respectively for each predictor derived from equation (6), for simplicity, we denote it as $p(\gamma_{kj} = 1) = \pi_{kj}$. The MCMC estimation of the parameter set ($\boldsymbol{\gamma_k}$,$\boldsymbol{\beta_k}$, σ_k^2) can be implemented using Gibbs sampling:

- sample each γ_{kj} from $p(\gamma_{kj}) = (1 - \pi_{kj})p(\beta_{kj}; 0, c\sigma_k^2)\delta_0 + \pi_{kj}p(\beta_{kj}; 0, c\sigma_k^2)I_{\gamma_{kj}=1}$

- sample $\boldsymbol{\beta_k}$ from $\mathcal{N}((\boldsymbol{X'X}/\sigma_k^2+\boldsymbol{D^{-1}})^{-1}\boldsymbol{X'y}/\sigma_k^2, (\boldsymbol{X'X}/\sigma_k^2+\boldsymbol{D^{-1}})^{-1})$ and $\boldsymbol{D} = diag(\sigma_k^2\gamma_{kj})$

- update π_{kj} based on equation (6) which involves other updated parameters.

B.1.2 Updating θ_{km} and τ_{km}

The θ_{km} and τ_{km} are sampled from their joint posterior conditional distribution. Considering the terms that only contain these two parameters from equation (2), the posterior full conditional of θ_{km} and τ_{km} can be formulated as:

$$\begin{aligned}\theta_{km}, \tau_{km}|rest \\ &\propto \prod_{j=1}^{p}\{C_k(\nu_j,\Theta)^{-1}exp(2\theta_{km}\gamma_{kj}\gamma_{mj})\}w^{\tau_{km}}(1-w)^{1-\tau_{km}}[(1-\tau_{km})\delta_0 \\ &+ \tau_{km}\frac{\beta_\theta^{\alpha_\theta}}{\Gamma(\alpha_\theta)}\theta_{km}^{\alpha_\theta-1}e^{-\beta_\theta\theta_{km}}]\end{aligned} \tag{B.2}$$

Considering the fact that the normalizing constant is not analytically tractable, we apply Metropolis-Hastings method to sample θ_{km} and τ_{km}. The MCMC sampling scheme in this step involves the approach described by Gottardo and Raftery (cite), which for each iteration, the entire step involves two steps, between-model move and within-model move. This approach is also called reversible jump Markov chain Monte Carlo (RJMCMC). Following shows the specific algorithm:

- between-model move

  - If the $\tau_{km} = 1$ in the current state, we take $\tau_{km}^* = 0$ and $\theta_{km}^* = 0$ as our proposed sample pairs.

  - If the $\tau_{km} = 0$ in the current state, we propose $\tau_{km}^* = 1$ and θ_{km}^* from the proposal distribution, we chose to set up the proposal distribution as $f(\theta_{km}^*) = Gamma(\alpha_\theta^*, \beta_\theta^*)$. The MH ratio given the posterior and proposal distributions

when moving from $\tau_{km} = 1$ to $\tau_{km} = 0$ can be written as

$$\begin{aligned} R_{1\to 0} &= \frac{p(\theta^*_{km}, \tau^*_{km}|rest)/f(\theta^*_{km})}{p(\theta_{km}, \tau_{km}|rest)/f(\theta_{km})} \\ &= \frac{1-w}{w} \cdot \frac{\Gamma(\alpha_\theta)(\beta^*_\theta)^{\alpha^*_\theta}}{\Gamma(\alpha^*_\theta)(\beta^*_\theta)^{\alpha_\theta}} (\theta_{km})^{\alpha^*_\theta, -\alpha_\theta} \\ &\cdot exp\{(\beta_\theta - \beta^*_\theta)\theta_{km}\} \cdot \prod_{j=1}^{P} \frac{C_k(\nu_j, \Theta) \cdot exp(-2\theta_{km}\gamma_{kj}\gamma_{mj})}{C_k(\nu_j, \Theta^*)} \end{aligned} \tag{B.3}$$

Where Θ^* denotes the updated Θ with element θ_{km} replaced by θ^*_{km}.

When moving from $\tau_{km} = 0$ to $\tau_{km} = 1$, the ratio can be expressed as

$$\begin{aligned} R_{0\to 1} &= \frac{p(\theta^*_{km}, \tau^*_{km}|rest)/f(\theta^*_{km})}{p(\theta_{km}, \tau_{km}|rest)/f(\theta_{km})} \\ &= \frac{w}{1-w} \cdot \frac{\Gamma(\alpha^*_\theta)(\beta_\theta)^{\alpha_\theta}}{\Gamma(\alpha_\theta)(\beta^*_\theta)^{\alpha^*_\theta}} (\theta^*_{km})^{\alpha_\theta, -\alpha^*_\theta} \\ &\cdot exp\{(\beta^*_\theta - \beta_\theta)\theta^*_{km}\} \cdot \prod_{j=1}^{P} \frac{C_k(\nu_j, \Theta) \cdot exp(2\theta^*_{km}\gamma_{kj}\gamma_{mj})}{C_k(\nu_j, \Theta^*)} \end{aligned} \tag{B.4}$$

- within-model move

  When τ_{km} sampled from the between-model move euqals 1, we propose within-model move. That is, we propose another θ^*_{km} from the same proposal distribution and compute the MH ratio as

$$\begin{aligned} R_{0\to 1} &= \frac{p(\theta^*_{km}, \tau^*_{km}|rest)/f(\theta^*_{km})}{p(\theta_{km}, \tau_{km}|rest)/f(\theta_{km})} \\ &= \Big\{\frac{\theta^*_{km}}{\theta_{km}}\Big\}^{\alpha_\theta - \alpha^*_\theta} \cdot exp\{(\beta^*_\theta - \beta_\theta)(\theta^*_{km} - \theta_{km})\} \cdot \prod_{j=1}^{P} \frac{C_k(\nu_j, \Theta) exp(2(\theta^*_{km} - \theta_{km})\gamma_{kj}\gamma_{mj})}{C_k(\nu_j, \Theta^*)} \end{aligned} \tag{B.5}$$

B.1.3 Updating ν_j

Given the prior of ν_j, the prosterior distribution is derived as

$$\nu_j|rest \propto \frac{exp(\nu_j(a + \mathbf{1}^T\boldsymbol{\gamma_j}))}{C(\nu_j, \boldsymbol{\Theta})(1 + e^{\nu_j})^{a+b}} \tag{B.6}$$

Similarly, the normalizing constant is not analytical tractable, we apply MH approach in this step as well. For each j, we propose q^* from proposal density, here we apply the same distribution, $Beta(a^*, b^*)$ distribution, and our new ν_j can be calculated from $\nu_j^* = logit(q^*)$. The MH ratio can be expressed as

$$\begin{aligned} R_\nu &= \frac{p(\nu_j^*|rest)q(\nu_j)}{p(\nu_j|rest)q(\nu_j^*)} \\ &= \frac{C_k(\nu_j, \Theta)}{C_k(\nu_j^*, \Theta)} \cdot \frac{exp((\nu_j^* - \nu_j)(\mathbf{1}^T\boldsymbol{\gamma_j} + a - a^*)) \cdot (1 + exp(\nu_j))^{a+b-a^*-b^*}}{(1 + exp(\nu_j^*))^{a+b-a^*-b^*}} \end{aligned} \tag{B.7}$$

B.2 Table of pathway gene membership and sample size information

Pathway	Genes					
Apoptosis	BAK1	BAX	BID	BCL2L11	CASP7	BAD
	BCL2	BCL2L1	BIRC2			
Breast reactive	CAV1	MYH11	RAB11A	RAB11B	CTNNB1	GAPDH
	RBM15					
Cell cycle	CDK1	CCNB1	CCNE1	CCNE2	CDKN1B	PCNA
	FOXM1					
Core reactive	CAV1	CTNNB1	RBM15	CDH1	CLDN7	
DNA damage response	TP53BP1	ATM	BRCA2	CHEK1	CHEK2	XRCC5
	MRE11A	TP53	RAD50	RAD51	XRCC1	
EMT	FN1	CDH2	COL6A1	CLDN7	CDH1	CTNNB1
	SERPINE1					
PI3K/AKT	AKT1	AKT2	AKT3	GSK3A	GSK3B	CDKN1B
	AKT1S1	TSC2	INPP4B	PTEN		
RAS/MAPK	ARAF	JUN	RAF1	MAPK8	MAPK1	MAPK3
	MAP2K1	MAPK14	RPS6KA1	YBX1		
RTK	EGFR	ERBB2	ERBB3	SHC1	SRC	
TSC/mTOR	EIF4EBP1	RPS6KB1	MTOR	RPS6	RB1	
Hormone receptor	ESR1	PGR	AR			
Hormone signaling (Breast)	INPP4B	GATA3	BCL2			

Table B.1: Pathway-Gene membership Table

	bar_rppa	bar_mRNA	bar_cna	bar_methy	bar_miRNA	bar_inter
ACC	46	79	90	80	80	45
BLCA	344	408	408	412	409	334
BRCA	869	1095	1080	785	756	498
CESC	171	304	295	307	307	160
CHOL	30	36	36	36	36	30
COAD	354	285	451	296	221	186
DLBC	33	48	48	48	47	33
ESCA	126	184	184	185	184	124
GBM	204	161	577	141	565	31
HNSC	346	520	522	528	486	304
KICH	63	66	66	66	66	63
KIRC	445	533	528	319	254	139
KIRP	207	290	288	275	291	203
LGG	423	516	513	516	512	416
LIHC	184	371	370	377	372	171
LUAD	362	515	516	458	450	309
LUSC	325	501	501	370	342	233
MESO	61	87	87	87	87	61
OV	402	305	582	10	475	0
PAAD	105	178	184	184	178	97
PCPG	79	179	162	179	179	77
PRAD	351	497	494	498	494	342
READ	129	94	165	98	90	68
SARC	218	259	257	261	259	211
SKCM	351	468	469	469	448	333
STAD	392	415	441	395	389	291
TGCT	114	150	150	150	150	114
THCA	368	505	502	507	506	362
THYM	90	120	123	124	124	87
UCEC	404	177	539	431	411	101
UCS	48	57	56	57	56	46
UVM	12	80	80	80	80	12

Table B.2: Number of the samples for each platform for pan-cancer; rppa is proteomic data, mRNA is gene expression, cna is copynumber alteration, methy is methylation, miRNA is microRNA, inter is the intersection of all platforms

Bibliography

[1] Jerry M Adams and Suzanne Cory. The bcl-2 arbiters of apoptosis and their growing role as cancer targets. *Cell death and differentiation*, 25(1):27, 2018.

[2] Hugo JWL Aerts, Emmanuel Rios Velazquez, Ralph TH Leijenaar, Chintan Parmar, Patrick Grossmann, Sara Carvalho, Johan Bussink, René Monshouwer, Benjamin Haibe-Kains, Derek Rietveld, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature communications*, 5, 2014.

[3] Leona S Aiken, Stephen G West, and Raymond R Reno. *Multiple regression: Testing and interpreting interactions*. Sage, 1991.

[4] Rehan Akbani, Patrick Kwok Shing Ng, Henrica MJ Werner, Maria Shahmoradgoli, Fan Zhang, Zhenlin Ju, Wenbin Liu, Ji-Yeon Yang, Kosuke Yoshihara, Jun Li, et al. A pan-cancer proteomic perspective on the cancer genome atlas. *Nature communications*, 5:3887, 2014.

[5] Dvir Aran, Sivan Sabato, and Asaf Hellman. Dna methylation of distal regulatory sites characterizes dysregulation of cancer genes. *Genome biology*, 14(3):R21, 2013.

[6] Artin Armagan, David B Dunson, and Jaeyong Lee. Generalized double pareto shrinkage. *Statistica Sinica*, 23(1):119, 2013.

[7] Shabnam Azadeh, Brian P Hobbs, Liangsuo Ma, David A Nielsen, F Gerard Moeller, and Veerabhadran Baladandayuthapani. Integrative bayesian analysis of neuroimaging-genetic data with application to cocaine dependence. *NeuroImage*, 125:813–824, 2016.

[8] David A Barbie, Pablo Tamayo, Jesse S Boehm, So Young Kim, Susan E Moody, Ian F Dunn, Anna C Schinzel, Peter Sandy, Etienne Meylan, Claudia Scholl, et al. Systematic rna interference reveals that oncogenic kras-driven cancers require tbk1. *Nature*, 462(7269):108, 2009.

[9] Maria Maddalena Barbieri and James O Berger. Optimal predictive model selection. *Annals of Statistics*, pages 870–897, 2004.

[10] Nematollah K Batmanghelich, Adrian V Dalca, Mert R Sabuncu, and Polina Golland. Joint modeling of imaging and genetics. In *Information Processing in Medical Imaging*, pages 766–777. Springer, 2013.

[11] Ashton C Berger, Anil Korkut, Rupa S Kanchi, Apurva M Hegde, Walter Lenoir, Wenbin Liu, Yuexin Liu, Huihui Fan, Hui Shen, Visweswaran Ravikumar, et al. A comprehensive pan-cancer molecular study of gynecologic and breast cancers. *Cancer Cell*, 2018.

[12] Matteo Bersanelli, Ettore Mosca, Daniel Remondini, Enrico Giampieri, Claudia Sala, Gastone Castellani, and Luciano Milanesi. Methods for the integration of multi-omics data: mathematical aspects. *BMC bioinformatics*, 17(2):S15, 2016.

[13] Anirban Bhattacharya, Debdeep Pati, Natesh S Pillai, and David B Dunson. Dirichlet–laplace priors for optimal shrinkage. *Journal of the American Statistical Association*, 110(512):1479–1490, 2015.

[14] Fonnet E Bleeker, Remco J Molenaar, and Sieger Leenstra. Recent advances in the molecular understanding of glioblastoma. *Journal of neuro-oncology*, 108(1):11–27, 2012.

[15] Maciej Bobowicz, Marcin Skrzypski, Piotr Czapiewski, Michał Marczyk, Agnieszka Maciejewska, Michał Jankowski, Anna Szulgo-Paczkowska, Wojciech Zegarski, Ryszard Pawłowski, Joanna Polańska, et al. Prognostic value of 5-microrna based signature in t2-t3n0 colon cancer. *Clinical & experimental metastasis*, 33(8):765–773, 2016.

[16] G Bowers, D Reardon, T Hewitt, P Dent, RB Mikkelsen, K Valerie, G Lammering, C Amir, and RK Schmidt-Ullrich. The relative role of erbb1–4 receptor tyrosine kinases in radiation signal transduction responses of human carcinoma cells. *Oncogene*, 20(11):1388, 2001.

[17] Ramesh Butti, Sumit Das, Vinoth Prasanna Gunasekaran, Amit Singh Yadav, Dhiraj Kumar, and Gopal C Kundu. Receptor tyrosine kinases (rtks) in breast cancer: signaling, therapeutic implications and challenges. *Molecular cancer*, 17(1):34, 2018.

[18] Joshua D Campbell, Christina Yau, Reanne Bowlby, Yuexin Liu, Kevin Brennan, Huihui Fan, Alison M Taylor, Chen Wang, Vonn Walter, Rehan Akbani, et al. Genomic, pathway network, and immunologic features distinguishing squamous carcinomas. *Cell reports*, 23(1):194, 2018.

[19] Carlos M Carvalho, Nicholas G Polson, and James G Scott. The horseshoe estimator for sparse signals. *Biometrika*, page asq017, 2010.

[20] G Castellano, L Bonilha, LM Li, and F Cendes. Texture analysis of medical images. *Clinical radiology*, 59(12):1061–1069, 2004.

[21] Ahmad Chaddad and Camel Tanougast. Extracted magnetic resonance texture features discriminate between phenotypes and are associated with overall survival in glioblastoma multiforme patients. *Medical & biological engineering & computing*, pages 1–12, 2016.

[22] Ami Citri and Yosef Yarden. Egf–erbb signalling: towards the systems level. *Nature reviews Molecular cell biology*, 7(7):505, 2006.

[23] Thibaud P Coroller, Patrick Grossmann, Ying Hou, Emmanuel Rios Velazquez, Ralph TH Leijenaar, Gretchen Hermann, Philippe Lambin, Benjamin Haibe-Kains, Raymond H Mak, and Hugo JWL Aerts. Ct-based radiomic signature predicts distant metastasis in lung adenocarcinoma. *Radiotherapy and Oncology*, 114(3):345–350, 2015.

[24] Anneleen Daemen, Olivier Gevaert, Fabian Ojeda, Annelies Debucquoy, Johan AK Suykens, Christine Sempoux, Jean-Pascal Machiels, Karin Haustermans, and Bart De Moor. A kernel-based integration of genome-wide data for clinical decision support. *Genome Medicine*, 1(4):39, 2009.

[25] Clare G Fedele, Lisa M Ooms, Miriel Ho, Jessica Vieusseux, Sandra A O'Toole, Ewan K Millar, Elena Lopez-Knowles, Absorn Sriratana, Rajendra Gurung, Laura Baglietto, et al. Inositol polyphosphate 4-phosphatase ii regulates pi3k/akt signaling and is lost in human basal-like breast cancers. *Proceedings of the National Academy of Sciences*, 107(51):22231–22236, 2010.

[26] E Melo Felipe De Sousa, Louis Vermeulen, Evelyn Fessler, and Jan Paul Medema. Cancer heterogeneity—a multifaceted view. *EMBO reports*, 14(8):686–695, 2013.

[27] Balaji Ganeshan, Sandra Abaleke, RC Young, Christopher R Chatwin, and Kenneth A Miles. Texture analysis of non-small cell lung cancer on unenhanced computed tomography: initial evidence for a relationship with tumour glucose metabolism and stage. *Cancer imaging*, 10(1):137–143, 2010.

[28] Balaji Ganeshan, Kenneth A Miles, Rupert CD Young, Christopher R Chatwin, Hugh MD Gurling, and Hugo D Critchley. Three-dimensional textural analysis of brain images reveals distributed grey-matter abnormalities in schizophrenia. *European radiology*, 20(4):941–948, 2010.

[29] Edward I George and Robert E McCulloch. Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993.

[30] John Geweke et al. *Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments*, volume 196. Federal Reserve Bank of Minneapolis, Research Department Minneapolis, MN, USA, 1991.

[31] Christina Gewinner, Zhigang C Wang, Andrea Richardson, Julie Teruya-Feldstein, Dariush Etemadmoghadam, David Bowtell, Jordi Barretina,

William M Lin, Lucia Rameh, Leonardo Salmena, et al. Evidence that inositol polyphosphate 4-phosphatase type ii is a tumor suppressor that inhibits pi3k signaling. *Cancer cell*, 16(2):115–125, 2009.

[32] Jim E Griffin, Philip J Brown, et al. Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis*, 5(1):171–188, 2010.

[33] Justin Guinney, Rodrigo Dienstmann, Xin Wang, Aurélien De Reyniès, Andreas Schlicker, Charlotte Soneson, Laetitia Marisa, Paul Roepman, Gift Nyamundanda, Paolo Angelino, et al. The consensus molecular subtypes of colorectal cancer. *Nature medicine*, 21(11):1350, 2015.

[34] David A Gutman, Lee AD Cooper, Scott N Hwang, Chad A Holder, JingJing Gao, Tarun D Aurora, William D Dunn Jr, Lisa Scarpace, Tom Mikkelsen, Rajan Jain, et al. Mr imaging predictors of molecular profile and survival: multi-institutional study of the tcga glioblastoma data set. *Radiology*, 267(2):560–569, 2013.

[35] Sonja Hänzelmann, Robert Castelo, and Justin Guinney. Gsva: gene set variation analysis for microarray and rna-seq data. *BMC bioinformatics*, 14(1):7, 2013.

[36] Robert M Haralick. Statistical and structural approaches to texture. *Proceedings of the IEEE*, 67(5):786–804, 1979.

[37] Robert M Haralick. Lg shapiro. computer and robot vision, vol. 1. 1992.

[38] Robert M Haralick, Karthikeyan Shanmugam, and Its' Hak Dinstein. Textural features for image classification. *Systems, Man and Cybernetics, IEEE Transactions on*, (6):610–621, 1973.

[39] Yehudit Hasin, Marcus Seldin, and Aldons Lusis. Multi-omics approaches to disease. *Genome biology*, 18(1):83, 2017.

[40] Trevor J Hastie and Robert J Tibshirani. *Generalized additive models*, volume 43. CRC Press, 1990.

[41] J Graeme Hodgson, Ru-Fang Yeh, Amrita Ray, Nicholas J Wang, Ivan Smirnov, Mamie Yu, Sujatmi Hariono, Joachim Silber, Heidi S Feiler, Joe W Gray, et al. Comparative analyses of gene copy number and mrna expression in glioblastoma multiforme tumors and xenografts. *Neuro-oncology*, 11(5):477–487, 2009.

[42] Myles C Hodgson, Elena I Deryugina, Egla Suarez, Sandra M Lopez, Dong Lin, Hui Xue, Ivan P Gorlov, Yuzhuo Wang, and Irina U Agoulnik. Inpp4b suppresses prostate cancer cell invasion. *Cell Communication and Signaling*, 12(1):61, 2014.

[43] Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[44] Frederik Holst, Phillip R Stahl, Christian Ruiz, Olaf Hellwinkel, Zeenath Jehan, Marc Wendland, Annette Lebeau, Luigi Terracciano, Khawla Al-Kuraya, Fritz Jänicke, et al. Estrogen receptor alpha (esr1) gene amplification is frequent in breast cancer. *Nature genetics*, 39(5):655, 2007.

[45] Leland S Hu, Shuluo Ning, Jennifer M Eschbacher, Leslie C Baxter, Nathan Gaw, Sara Ranjbar, Jonathan Plasencia, Amylou C Dueck, Sen Peng, Kris A Smith, et al. Radiogenomics to characterize regional genetic heterogeneity in glioblastoma. *Neuro-Oncology*, 19(1):128–137, 2017.

[46] Clifford A Hudis. Trastuzumab—mechanism of action and use in clinical practice. *New England Journal of Medicine*, 357(1):39–51, 2007.

[47] J Larry Jameson and Dan L Longo. Precision medicine—personalized, problematic, and promising. *Obstetrical & Gynecological Survey*, 70(10):612–614, 2015.

[48] Elizabeth M Jennings, Jeffrey S Morris, Raymond J Carroll, Ganiraju C Manyam, and Veerabhadran Baladandayuthapani. Hierarchical bayesian methods for integration of various types of genomics data. In *Genomic Signal Processing and Statistics,(GENSIPS), 2012 IEEE International Workshop on*, pages 5–8. IEEE, 2012.

[49] Elizabeth M Jennings, Jeffrey S Morris, Raymond J Carroll, Ganiraju C Manyam, and Veerabhadran Baladandayuthapani. Bayesian methods for expression-based integration of various types of genomics data. *EURASIP Journal on Bioinformatics and Systems Biology*, 2013(1):13, 2013.

[50] Elizabeth M Jennings, Jeffrey S Morris, Ganiraju C Manyam, Raymond J Carroll, and Veerabhadran Baladandayuthapani. Bayesian models for flexible integrative analysis of multi-platform genomics data. *Book: Integrating omics data: statistical and computational methods. Cambridge University Press;1 edition*, 2015.

[51] Valen E Johnson and David Rossell. On the use of non-local prior densities in bayesian hypothesis tests. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(2):143–170, 2010.

[52] Iain M Johnstone and Bernard W Silverman. Needles and straw in haystacks: Empirical bayes estimates of possibly sparse sequences. *Annals of Statistics*, pages 1594–1649, 2004.

[53] Peter A Jones. Functions of dna methylation: islands, start sites, gene bodies and beyond. *Nature Reviews Genetics*, 13(7):484, 2012.

[54] Purvesh Khatri, Marina Sirota, and Atul J Butte. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS computational biology*, 8(2):e1002375, 2012.

[55] Philipp Kickingereder, Sina Burth, Antje Wick, Michael Götz, Oliver Eidel, Heinz-Peter Schlemmer, Klaus H Maier-Hein, Wolfgang Wick, Martin Bendszus, Alexander Radbruch, et al. Radiomic profiling of glioblastoma: identifying an imaging predictor of patient survival with improved performance over established clinical and radiologic risk models. *Radiology*, 280(3):880–889, 2016.

[56] Young-Ho Kim, Joel Lachuer, Michel Mittelbronn, Werner Paulus, Benjamin Brokinkel, Kathy Keyvani, Ulrich Sure, Karsten Wrede, Sumihito Nobusawa, Yoichi Nakazato, et al. Alterations in the rb1 pathway in low-grade diffuse gliomas lacking common genetic alterations. *Brain Pathology*, 21(6):645–651, 2011.

[57] Lidija Klampfer. The role of signal transducers and activators of transcription in colon cancer. *Front Biosci*, 13(3):2888–2899, 2008.

[58] Philippe Lambin, Emmanuel Rios-Velazquez, Ralph Leijenaar, Sara Carvalho, Ruud GPM van Stiphout, Patrick Granton, Catharina ML Zegers, Robert Gillies, Ronald Boellard, André Dekker, et al. Radiomics: extracting more information from medical images using advanced feature analysis. *European journal of cancer*, 48(4):441–446, 2012.

[59] Gert RG Lanckriet, Tijl De Bie, Nello Cristianini, Michael I Jordan, and William Stafford Noble. A statistical framework for genomic data fusion. *Bioinformatics*, 20(16):2626–2635, 2004.

[60] I Lawrence and Kuei Lin. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, pages 255–268, 1989.

[61] A Lebeau, TJ Grob, F Holst, N Seyedi-Fazlollahi, H Moch, Luigi Terracciano, A Turzynski, M Choschzick, G Sauter, and R Simon. Oestrogen receptor gene (esr1) amplification is frequent in endometrial carcinoma and its precursor lesions. *The Journal of pathology*, 216(2):151–157, 2008.

[62] Eunjung Lee, Han-Yu Chuang, Jong-Won Kim, Trey Ideker, and Doheon Lee. Inferring pathway activity toward precise disease classification. *PLoS computational biology*, 4(11):e1000217, 2008.

[63] HJ Lee, AN Seo, EJ Kim, MH Jang, YJ Kim, JH Kim, SW Kim, HS Ryu, IA Park, SA Im, et al. Prognostic and predictive values of egfr overexpression and egfr copy number alteration in her2-positive breast cancer. *British journal of cancer*, 112(1):103, 2015.

[64] J Lee, R Jain, K Khalil, B Griffith, R Bosca, G Rao, and A Rao. Texture feature ratios from relative cbv maps of perfusion mri are associated with patient survival in glioblastoma. *American Journal of Neuroradiology*, 37(1):37–43, 2016.

[65] Fan Li and Nancy R Zhang. Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics. *Journal of the American statistical association*, 105(491):1202–1214, 2010.

[66] Jun Li, Yiling Lu, Rehan Akbani, Zhenlin Ju, Paul L Roebuck, Wenbin Liu, Ji-Yeon Yang, Bradley M Broom, Roeland GW Verhaak, David W Kane, et al. Tcpa: a resource for cancer functional proteomics data. *Nature methods*, 10(11):1046, 2013.

[67] Arthur Liberzon, Aravind Subramanian, Reid Pinchback, Helga Thorvaldsdóttir, Pablo Tamayo, and Jill P Mesirov. Molecular signatures database (msigdb) 3.0. *Bioinformatics*, 27(12):1739–1740, 2011.

[68] Haiyan Liu, Jianhui Shi, Myra L Wilkerson, and Fan Lin. Immunohistochemical evaluation of gata3 expression in tumors and normal tissuesa useful immunomarker for breast and urothelial carcinomas. *American journal of clinical pathology*, 138(1):57–64, 2012.

[69] Ganiraju Manyam, Cristina Ivan, George A Calin, and Kevin R Coombes. targethub: a programmable interface for mirna–gene interactions. *Bioinformatics*, 29(20):2657–2658, 2013.

[70] Matthew J McAuliffe, Francois M Lalonde, Delia McGarry, William Gandler, Karl Csaky, and Benes L Trus. Medical image processing, analysis and visualization in clinical research. In *Computer-Based Medical Systems, 2001. CBMS 2001. Proceedings. 14th IEEE Symposium on*, pages 381–386. IEEE, 2001.

[71] Carroll RJ Manyan G Baladandayuthapani V McGuffey et al., Morris JS. pibag: Pathway-based integrative bayesian modeling of multiplatform genomics data. Under Review, 2018.

[72] Rohit Mehra, Sooryanarayana Varambally, Lei Ding, Ronglai Shen, Michael S Sabel, Debashis Ghosh, Arul M Chinnaiyan, and Celina G Kleer. Identification of gata3 as a breast cancer prognostic marker by global gene expression meta-analysis. *Cancer research*, 65(24):11259–11264, 2005.

[73] Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.

[74] Malgorzata Milewska, Mattia Cremona, Clare Morgan, John O'Shea, Aoife Carr, Sri H Velanki, Ann M Hopkins, Sinead Toomey, Stephen F Madden, Bryan T Hennessy, et al. Development of a personalized therapeutic strategy for erbb-gene-mutated cancers. *Therapeutic Advances in Medical Oncology*, 10:1758834017746040, 2018.

[75] Toby J Mitchell and John J Beauchamp. Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):1023–1032, 1988.

[76] Nicola Montano, Tonia Cenci, Maurizio Martini, Quintino Giorgio D'Alessandris, Federica Pelacchi, Lucia Ricci-Vitiani, Giulio Maira, Ruggero De Maria, Luigi Maria Larocca, and Roberto Pallini. Expression of egfrviii in glioblastoma: prognostic significance revisited. *Neoplasia*, 13(12):1113–IN6, 2011.

[77] Mitsutoshi Nakamura, Noboru Konishi, Shigeru Tsunoda, Yoshio Hiasa, Toshihide Tsuzuki, Takuo Inui, and Toshisuke Sakaki. Retinoblastoma protein expression and mib-1 correlate with survival of patients with malignant astrocytoma. *Cancer*, 80(2):242–249, 1997.

[78] Mitsutoshi Nakamura, Yasuhiro Yonekawa, Paul Kleihues, and Hiroko Ohgaki. Promoter hypermethylation of the rb1 gene in glioblastomas. *Laboratory investigation*, 81(1):77–82, 2001.

[79] Dougu Nam and Seon-Young Kim. Gene-set approach for expression pattern analysis. *Briefings in bioinformatics*, 9(3):189–197, 2008.

[80] Inga Nazarenko, Sanna-Maria Hede, Xiaobing He, Anna Hedrén, James Thompson, Mikael S Lindström, and Monica Nistér. Pdgf and pdgf receptors in glioma. *Upsala journal of medical sciences*, 117(2):99–112, 2012.

[81] Cancer Genome Atlas Network et al. Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature*, 517(7536):576, 2015.

[82] Cancer Genome Atlas Research Network et al. Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, 489(7417):519, 2012.

[83] Cancer Genome Atlas Research Network et al. Integrated genomic and molecular characterization of cervical cancer. *Nature*, 543(7645):378, 2017.

[84] Cancer Genome Atlas (TCGA) Research Network et al. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(7216):1061, 2008.

[85] Manal Nicolasjilwan, Ying Hu, Chunhua Yan, Daoud Meerzaman, Chad A Holder, David Gutman, Rajan Jain, Rivka Colen, Daniel L Rubin, Pascal O Zinn, et al. Addition of mr imaging features and genetic biomarkers strengthens glioblastoma survival prediction in tcga patients. *Journal of Neuroradiology*, 42(4):212–221, 2015.

[86] Vesna Nikolov, Miodrag Stojanovic, Aleksandar Kostic, Misa Radisavljevic, Natasa Simonovic, Boban Jelenkovic, and Luka Berilazic. Factors affecting the survival of patients with glioblastoma multiforme. *month*, 38:56–70, 2018.

[87] Rolando J Olivares, Akhila Rao, Ganeswara Rao, Jeffrey S Morris, and Veerabhadran Baladandayuthapani. Integrative analysis of multi-modal correlated imaging-genomics data in glioblastoma. In *Genomic Signal Processing*

and Statistics (GENSIPS), 2013 IEEE International Workshop on, pages 5–8. IEEE, 2013.

[88] Trevor Park and George Casella. The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

[89] Rolando Perez-Lorenzo, Kamraan Z Gill, Che-Hung Shen, Feng X Zhao, Bin Zheng, Hans-Joachim Schulze, David N Silvers, Georg Brunner, and Basil A Horst. A tumor suppressor function for the lipid phosphatase inpp4b in melanocytic neoplasms. *Journal of Investigative Dermatology*, 134(5):1359–1368, 2014.

[90] Christine Peterson, Francesco C Stingo, and Marina Vannucci. Bayesian inference of multiple gaussian graphical models. *Journal of the American Statistical Association*, 110(509):159–174, 2015.

[91] Binh Phan, Shahana Majid, Sarah Ursu, David de Semir, Mehdi Nosrati, Vladimir Bezrookove, Mohammed Kashani-Sabet, and Altaf A Dar. Tumor suppressor role of microrna-1296 in triple-negative breast cancer. *Oncotarget*, 7(15):19519, 2016.

[92] Daniel L Roden, Laura A Baker, Benjamin Elsworth, Chia-Ling Chan, Kate Harvey, Niantao Deng, Sunny Wu, Aurelie Cazet, Radhika Nair, and Alexander Swarbrick. Single cell transcriptomics reveals molecular subtype and functional heterogeneity in models of breast cancer. *bioRxiv*, page 282079, 2018.

[93] Yuanbin Ru, Katerina J Kechris, Boris Tabakoff, Paula Hoffman, Richard A Radcliffe, Russell Bowler, Spencer Mahaffey, Simona Rossi, George A Calin, Lynne Bemis, et al. The multimir r package and database: integration of microrna–target interactions along with their disease and drug associations. *Nucleic acids research*, 42(17):e133–e133, 2014.

[94] Aaron M Rutman and Michael D Kuo. Radiogenomics: creating a link between molecular diagnostics and diagnostic imaging. *European journal of radiology*, 70(2):232–241, 2009.

[95] Anthony J Sadler and Bryan RG Williams. Interferon-inducible antiviral effectors. *Nature reviews immunology*, 8(7):nri2314, 2008.

[96] L Salmena, P Shaw, I Fans, B Rosen, H Risch, C Mitchell, P Sun, SA Narod, J Kotsopoulos, et al. Prognostic value of inpp4b protein immunohistochemistry in ovarian cancer. *European journal of gynaecological oncology*, 36(3):260–267, 2015.

[97] Aaron J Schetter, Giang Huong Nguyen, Elise D Bowman, Ewy A Mathé, Siu Tsan Yuen, Jason E Hawkes, Carlo M Croce, Suet Yi Leung, and Curtis C Harris. Association of inflammation-related and microrna gene expression with cancer-specific mortality of colon adenocarcinoma. *Clinical Cancer Research*, pages 1078–0432, 2009.

[98] James G Scott and James O Berger. Bayes and empirical-bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics*, pages 2587–2619, 2010.

[99] Devinderjit Sivia and John Skilling. *Data analysis: a Bayesian tutorial*. OUP Oxford, 2006.

[100] Martha L Slattery, Abbie Lundgreen, Kristina L Bondurant, and Roger K Wolff. Interferon-signaling pathway: associations with colon and rectal cancer risk and subsequent survival. *Carcinogenesis*, 32(11):1660–1667, 2011.

[101] David A Solomon, Jung-Sik Kim, Sultan Jenkins, Habtom Ressom, Michael Huang, Nicholas Coppa, Lauren Mabanta, Darell Bigner, Hai Yan, Walter Jean, et al. Identification of p18ink4c as a tumor suppressor gene in glioblastoma multiforme. *Cancer research*, 68(8):2564–2569, 2008.

[102] Francesco C Stingo, Michele Guindani, Marina Vannucci, and Vince D Calhoun. An integrative bayesian modeling approach to imaging genetics. *Journal of the American Statistical Association*, 108(503):876–891, 2013.

[103] Roger Stupp, Monika E Hegi, Warren P Mason, Martin J van den Bent, Martin JB Taphoorn, Robert C Janzer, Samuel K Ludwin, Anouk Allgeier, Barbara Fisher, Karl Belanger, et al. Effects of radiotherapy with concomitant and adjuvant temozolomide versus radiotherapy alone on survival in glioblastoma in a randomised phase iii study: 5-year analysis of the eortc-ncic trial. *The lancet oncology*, 10(5):459–466, 2009.

[104] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.

[105] Jessica D Tenenbaum, Michael G Walker, Paul J Utz, and Atul J Butte. Expression-based pathway signature analysis (epsa): Mining publicly available microarray data for insight into human disease. *BMC medical genomics*, 1(1):51, 2008.

[106] Lu Tian, Steven A Greenberg, Sek Won Kong, Josiah Altschuler, Isaac S Kohane, and Peter J Park. Discovering statistically significant pathways in expression profiling studies. *Proceedings of the National Academy of Sciences*, 102(38):13544–13549, 2005.

[107] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[108] Katarzyna Tomczak, Patrycja Czerwińska, and Maciej Wiznerowicz. The cancer genome atlas (tcga): an immeasurable source of knowledge. *Contemporary oncology*, 19(1A):A68, 2015.

[109] John Tomfohr, Jun Lu, and Thomas B Kepler. Pathway level analysis of gene expression using singular value decomposition. *BMC bioinformatics*, 6(1):225, 2005.

[110] Roel GW Verhaak, Katherine A Hoadley, Elizabeth Purdom, Victoria Wang, Yuan Qi, Matthew D Wilkerson, C Ryan Miller, Li Ding, Todd Golub, Jill P Mesirov, et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in pdgfra, idh1, egfr, and nf1. *Cancer cell*, 17(1):98–110, 2010.

[111] Hao Wang. Sparse seemingly unrelated regression modelling: Applications in finance and econometrics. *Computational Statistics & Data Analysis*, 54(11):2866–2877, 2010.

[112] Wenting Wang, Veerabhadran Baladandayuthapani, Jeffrey S Morris, Bradley M Broom, Ganiraju Manyam, and Kim-Anh Do. ibag: integrative bayesian analysis of high-dimensional multiplatform genomics data. *Bioinformatics*, 29(2):149–159, 2013.

[113] LJ Wei. The accelerated failure time model: a useful alternative to the cox regression model in survival analysis. *Statistics in medicine*, 11(14-15):1871–1879, 1992.

[114] John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Mills Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, Joshua M Stuart, Cancer Genome Atlas Research Network, et al. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113, 2013.

[115] Thomas F Westbrook, Eric S Martin, Michael R Schlabach, Yumei Leng, Anthony C Liang, Bin Feng, Jean J Zhao, Thomas M Roberts, Gail Mandel, Gregory J Hannon, et al. A genetic screen for candidate tumor suppressors identifies rest. *Cell*, 121(6):837–848, 2005.

[116] Gary Wilk and Rosemary Braun. Integrative analysis reveals disrupted pathways regulated by micrornas in cancer. *Nucleic acids research*, 46(3):1089–1101, 2017.

[117] Simon N Wood. mgcv: Gams and generalized ridge regression for r. *R news*, 1(2):20–25, 2001.

[118] Dong Yin, Seishi Ogawa, Norihiko Kawamata, Patrizia Tunici, Gaetano Finocchiaro, Marica Eoli, Christian Ruckert, Thien Huynh, Gentao Liu, Motohiro Kato, et al. High-resolution genomic copy number profiling of glioblastoma multiforme by single nucleotide polymorphism dna microarray. *Molecular Cancer Research*, 7(5):665–677, 2009.

[119] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

[120] Arnold Zellner. An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American statistical Association*, 57(298):348–368, 1962.

[121] Yiqun Zhang, Patrick Kwok-Shing Ng, Melanie Kucherlapati, Fengju Chen, Yuexin Liu, Yiu Huen Tsang, Guillermo de Velasco, Kang Jin Jeong, Rehan Akbani, Angela Hadjipanayis, et al. A pan-cancer proteogenomic atlas of pi3k/akt/mtor pathway alterations. *Cancer Cell*, 31(6):820–832, 2017.

[122] Mu Zhou, Lawrence Hall, Dmitry Goldgof, Robin Russo, Yoganand Balagurunathan, Robert Gillies, and Robert Gatenby. Radiologically defined ecological dynamics and clinical outcomes in glioblastoma multiforme: preliminary results. *Translational oncology*, 7(1):5–13, 2014.

[123] Yitan Zhu, Peng Qiu, and Yuan Ji. Tcga-assembler: open-source software for retrieving and processing tcga data. *Nature methods*, 11(6):599–600, 2014.

[124] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

[125] Hui Zou, Trevor Hastie, and Robert Tibshirani. Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286, 2006.

VITA

Youyi Zhang was born in Chifeng, Inner Mongolia, China, the daughter of Yuehao Zhang and Yanming Huang. When she was 5 years old, her family moved to Dalian, Liaoning, China. After completing her work at Dalian Yuming Senior High School, she entered Dongbei University of Finance and Economics (DUFE). She received the degree of Bachelor of Science with a major in Mathematics and Applied Mathematics (Economics oriented) in July 2012. For the next two years, she entered The University of California, San Diego, and obtained Master of Science degree with major in Statistics. In August of 2014, she entered the Ph.D. program in The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences majoring in Biostatistics.